TECHNISCHE UNIVERSITÄT CHEMNITZ

Informatik

Juniorprofessur Visual Computing

Masterarbeit

Intuitive Visualisierung universitätsinterner Publikationsdaten zur Unterstützung von Entscheidungsprozessen

Fabian Bolte

Chemnitz, den 22. September 2016

Prüfer: Jun.-Prof. Dr. Paul Rosenthal
Betreuer: Dipl.-Inf. Michael Heidt

Bolte, Fabian
Intuitive Visualisierung universitätsinterner Publikationsdaten zur Unterstützung von Entscheidungsprozessen
Masterarbeit, Informatik
Technische Universität Chemnitz, September 2016

Abstract

Die vorliegende Arbeit nutzt die Publikationsdaten der TU Chemnitz zur Darstellung der Entwicklung von Kooperationen zwischen Instituten und Fakultäten über die Zeit. Dabei wird die Unzulänglichkeit gängiger Netzwerkanalysen mithilfe von Graphen, die komplexen Beziehungen um eine zeitliche Dimension zu erweitern, aufgezeigt. Stattdessen wird eine Anwendung auf Basis des Streamgraphen vorgestellt, welche nicht nur den Vergleich der Entwicklung beliebiger Kombinationen von Instituten und Fakultäten ermöglicht, sondern auch spezifische Auskünfte zu den Kooperationsarten und deren zeitlicher Verlagerung gibt. Dafür werden zwei Erweiterungen für den Streamgraphen vorgestellt, welche seinen Informationsumfang erweitern und ihn damit zur Erfüllung der gesetzten Anforderungen befähigen.

Inhaltsverzeichnis

1. Einleitung

In diesem Kapitel werden die Anforderungen an die Visualisierung definiert und die Struktur der folgenden Kapitel beschrieben.

1.1. Motivation

Visualisierungen bieten einem Nutzer schnelle Einblicke in ansonsten nur schwer zu erfassende Datensätze. Betrachtet man die Publikationen einer Universität, so ist die absolute Zahl von Veröffentlichungen schnell herauszufinden und auch die Zahlen der vergangenen Jahre können in textueller Form verglichen werden. Möchte man jedoch die Anzahl der Publikationen eines jeden Instituts dieser Universität aufzeigen, ihre Entwicklung verfolgen und ihre Anteile an den Gesamtpublikationen erfassen, hilft eine geeignete Visualisierung dabei, in kürzester Zeit einen guten Überblick über die allgemeine Lage zu gewinnen. Betrachtet man zusätzlich die Zusammenarbeiten einzelner Institute und ist an den Beziehungen zwischen den Fakultäten interessiert, so wird die Aufgabe umso komplexer und fordert umso flexiblere Darstellungsformen.

1.2. Zielstellung

Die hier definierten Fragestellungen dienen einerseits zur Beurteilung der in Kapitel 3 untersuchten Darstellungen und werden andererseits in Kapitel 5 zur Evaluierung der in Kapitel 4 erstellten Anwendung genutzt.

Die Publikationen des Instituts einer Universität lassen sich in der Betrachtung von Kooperationen in vier Kategorien gliedern, je nachdem, welchen Instituten die beteiligten Autoren angehören. Die jeweilige Veröffentlichung entstand entweder innerhalb des Instituts, innerhalb der Fakultät (mit Koautoren anderer Institute), mit Instituten einer anderer Fakultät, oder mit externen Mitarbeitern. Eine Publikation kann auch mehreren dieser Kategorien angehören, indem sie beispielsweise je einen Koautor eines Instituts derselben und einer anderen Fakultät, sowie einen externen Mitarbeiter hatte. Die Gruppierungen dienen als Unterscheidungskriterium für die Kooperationen, an denen ein Institut beteiligt war.

Aus der entstehenden Visualisierung sollte nicht nur deutlich hervorgehen, welches Institut am meisten, beziehungsweise am wenigsten, kooperierte, sondern auch, wie sich diese Kooperationen auf die vier genannten Kategorien verteilen. Genauso sollte der Spitzenreiter einer jeden Kategorie festgestellt werden können.

Da in dieser Arbeit eine Anwendung entstehen soll, welche sowohl einen allgemeinen Überblick über die Kooperationslandschaft gibt, als auch spezifische Vergleiche der vorhandenen Institute zulässt, muss die Darstellung sowohl die Gegenüberstellung ausgewählter Institute, den Anteil des Instituts an seiner Fakultät, den Vergleich eines Instituts mit anderen Fakultäten, sowie den Vergleich mehrerer Fakultäten miteinander ermöglichen. Die Daten sollten dabei sowohl absolut, als auch prozentual dargestellt werden können.

Wird der zeitliche Aspekt eingebracht, um die Entwicklung der Daten aufzuzeigen, so ergeben sich auch weiterführende Fragestellungen. So kann beispielsweise jede oben genannte Frage nach dem maximalen, beziehungsweise minimalen, Wert in jeder Kategorie zusätzlich nach Jahren untergliedert werden. Weiterhin sollte ersichtlich sein, ob sich ein Institut in der jeweiligen Kategorie, oder insgesamt, gesteigert, oder verschlechtert hat und es sollte feststellbar sein, wie die Entwicklung eines gegebenen Instituts im Vergleich zu anderen Instituten ausfiel. Wenn beispielsweise eine Steigerung einer Fakultät über die Jahre sichtbar wird ist ebenfalls interessant, welches Institut an der Fakultät den größten Beitrag dazu leistete.

1.3. Gliederung

Nachdem bereits die Zielstellungen dieser Arbeit vorgestellt wurden, werden in Kapitel 2 die zur Verfügung gestellten Daten analysiert und die Methodik zur Extraktion der für unsere Darstellung wichtigen Daten vorgestellt. Dabei wird insbesondere darauf hingewiesen, wieso verschiedene, vorhandene Daten nicht zur Weiterverarbeitung verwendet werden können. In Kapitel 3 werden vorhandene Visualisierungen vorgestellt, welche bereits für ähnliche Daten verwendet wurden und auf ihre Vor- und Nachteile hin überprüft. Dabei werden einige Ansätze testweise implementiert, um ihre Verwendbarkeit für unseren speziellen Datensatz zu prüfen. Weiterhin wird auch die Visualisierung vorgestellt, welche schließlich als Grundlage für die in Kapitel 4 vorgestellte Implementierung der Anwendung genutzt wird. Die Vorstellung der Anwendung umfasst auch die Beschreibung all ihrer Interaktionsmöglichkeiten. Das Vorgehen und die Ergebnisse der Evaluierung werden in Kapitel 5 geschildert, wobei insbesondere auf Verbesserungsvorschläge und neue Funktionalitäten hingewiesen wird, welche der Anwendung zukünftig hinzugefügt werden sollten. Kapitel 6 zeigt einige Anwendungsbeispiele, welche das Potenzial der entstandenen Anwendung untermauern und zukünftigen Nutzern als Grundlage zu dessen Bedienung dienen soll. Die Zusammenfassung der hier entstandenen Arbeit folgt schließlich in Kapitel 7.

2. Daten

2.1. Einleitung

Daten sind die Grundlage einer jeden aussagekräftigen Visualisierung, weshalb in diesem Kapitel auf die Verfügbarkeit und Verarbeitung der Publikationsdaten der TU Chemnitz eingegangen wird. Dafür wird zuerst eine Analyse der gegebenen Datenbank, ihrer Relationen und Attribute durchgeführt, anschließend auf problematische Auffälligkeiten in dieser hingewiesen und schließlich die Extraktion der für die Visualisierung verwendbaren Daten beschrieben.

2.2. Datenbank

Die für die Visualisierung genutzten Daten werden aus einer Datenbank gelesen, welche auf einem Publikationsserver mit der Software Opus in Version 3.0 läuft. [Stu06] Dieser basiert auf einem Apache-Server mit einer MySQL-Datenbank, PHP in Version 4 oder höher, sowie curl. Es ist darauf zu achten, dass die tatsächliche Implementierung zum Teil von der Dokumentation abweicht, was sich unter anderem darin äußert, dass Attribute in Relationen der Datenbank hinzugefügt oder entfernt wurden.

In der Datenbank sind als wichtige Relationen opus, opus_autor, opus_diss, opus_inst, faculty_de, institute_de und resource_type_de zu nennen, welche im Datenbankschema 2.1 dargestellt und im Folgenden genauer beschrieben werden.

Als *Opus* wird im genutzten Framework jede in die Datenbank eingetragene Publikation bezeichnet und entsprechend in der Relation *opus* abgespeichert. Neben Titel, Dokumentart und Erscheinungsjahr der Veröffentlichung können weitere Attribute wie die Sprache, beteiligte Unternehmen, ISBN, und viele mehr angegeben werden. Des Weiteren wird jedem Dokument eine eindeutige Identifikationsnummer *source_opus* zugewiesen, welche zur besseren Verständlichkeit nachfolgend als OpusId bezeichnet wird.

Da eine Publikation von mehreren Autoren stammen kann, wird die Zuordnung der Veröffentlichungen zu ihren Autoren in eine weitere Relation *opus_autor* ausgelagert. In dieser wird neben der OpusId auch der Urheber der Arbeit und die Position seiner Nennung auf der Veröffentlichung (z.B. „1" für den Erstautor) gespeichert. Zusätzlich zu den in [Stu06] dokumentierten Attributen wurden die Kostenstelle des Autors, Informationen zur Kostenstelle, die Kosten der Publikation für diesen Autor und die

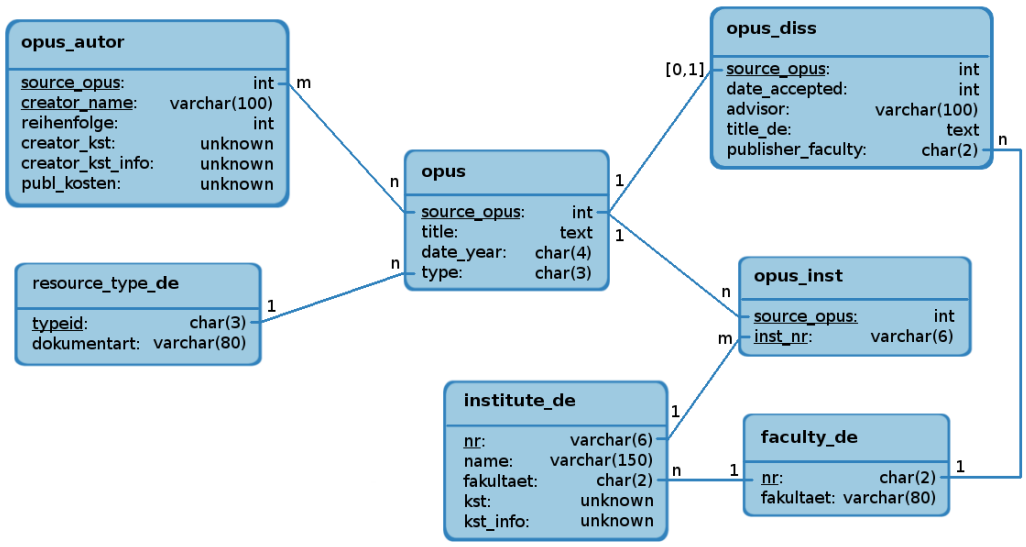


Abbildung 2.1.: Vereinfachtes Datenbankschema des Publikationsservers. Die wichtigsten Relationen der von Opus genutzten Datenbank sind opus, opus_autor, opus_diss, opus_inst, faculty_de, institute_de und resource_type_de. In dem hier gezeigten Schema werden nur die für die weitere Betrachtung bedeutungsvollen Attribute dargestellt. Weisen diese einen unbekannten Datentyp auf, so weichen diese von der Standardimplementierung von Opus ab und sind aus diesem Grund nicht dokumentiert.

Mittelherkunft hinzugefügt. Da diese Attribute in der Standardimplementierung nicht vorhanden sind und kein Zugriff auf das Informationsschema der Datenbank bestand, sind die Datentypen dieser hinzugefügten Attribute unbekannt. Die Zuordnung der Autoren zur Veröffentlichung erfolgt über die Verwendung derselben OpusId. Hat eine Veröffentlichung zum Beispiel fünf Autoren, so gibt es entsprechend fünf Einträge in der Relation *opus_autor*, welche dieselbe OpusId aufweisen. Dabei kann jeder Autor an beliebig vielen Veröffentlichungen mitgeschrieben haben.

Die Relation *resource_type_de* führt unter einer *typeid* 16 Dokumentarten auf, welche im Folgenden zusammen mit der Anzahl ihres Vorkommens am 22.09.2016 genannt werden:

- 1637 referierte Artikel in Fachzeitschriften
- 5834 nicht referierte Artikel in Fachzeitschriften

- 2737 referierte Konferenzbeiträge
- 1888 nicht referierte Konferenzbeiträge
- 1661 Konferenzabstracts
- 422 Bücher (Autorenschaft)
- 570 Bücher (Herausgeberschaft)
- 2518 Buchbeiträge
- 171 Zeitschriften/Buchreihen (Herausgeberschaft)
- 358 referierte, weitere Publikationen (u.a. Rezension, Lexikonbeitrag)
- 1219 nicht referierte, weitere Publikationen (u.a. Rezension, Lexikonbeitrag)
- 1139 Dissertationen
- 36 Habilitationen
- 158 Patente (Offenlegungsschrift, Patentschrift, Gebrauchsmusterschrift)
- 31 Poster
- 38 Preprints

Jede Publikation wird über die Verwendung des Fremdschlüssels *typeid* im Attribut *type* genau einer dieser Dokumentarten zugeordnet und jede Dokumentart kann beliebig vielen Publikationen zugeordnet sein.

Handelt es sich bei einer Veröffentlichung um eine Dissertation, so gibt es für diese einen Eintrag in der Relation *opus_diss*, in welcher zusätzliche Attribute wie das Datum der mündlichen Prüfung, der Name des ersten Betreuers, der Titel der Arbeit und die Fakultät, an welcher die Dissertation eingereicht wurde, gespeichert werden. Die Zuordnung erfolgt über die Verwendung derselben OpusId, wobei für jede Veröffentlichung maximal ein Eintrag in *opus_diss* zu finden ist.

Jede Arbeit wird in der Verknüpfungsrelation *opus_inst* mindestens einem Institut (zum Beispiel einer Professur) zugewiesen, indem ein Eintrag mit den entsprechenden Fremdschlüsseln der OpusId und der Identifikationsnummer eines Instituts (*inst_nr*) existiert. Sollten mehrere Institute an derselben Veröffentlichung beteiligt gewesen sein, so gibt es pro Institut einen Eintrag für diese OpusId. An dieser Stelle sei darauf hingewiesen, dass die Dokumentation [Stu06] die Beziehung zwischen *opus* und *opus_inst* als n-m-Beziehung beschreibt, obwohl diese, wie im Datenbankschema 2.1 zu erkennen ist, als Verknüpfungsrelation eine 1-n-Beziehung aufweisen muss und

erst mit einer weiteren 1-m-Beziehung zu *institute_de* die n-m-Beziehung zwischen *opus* und *institute_de* ensteht.

Die Institute werden in der Relation *institute_de* genauer beschrieben, indem jeder Identifikationsnummer ein Name, eine Fakultät und, erneut von der Dokumentation abweichend, eine Kostenstelle und Informationen zu dieser hinzugefügt werden. Die Datentypen der Kostenstelle und ihrer Informationen sind wie bei der Relation *opus_autor* unbekannt. Das Attribut der zugehörigen Fakultät ist selbst wieder ein Fremdschlüssel, welcher auf die Relation *faculty_de* verweist und dort eine genauere Bezeichnung erhält.

2.3. Aktuelle Zahlen

Am 22. September 2016 befinden sich in der genutzten Datenbank 9 Fakultäten, welche in 274 Institute unterteilt werden. 20417 Publikationen wurden abgespeichert, wovon 1175 Dissertationen sind. Die Relation *opus_autor* enthält 65089 Einträge, was bedeutet, dass für jede Veröffentlichung im Schnitt 3,18 Autoren eingetragen wurden.

Abbildung 2.2 zeigt, die Verteilung der Anzahl von Autoren pro Publikation für die drei referierten Dokumentarten. Im Gegensatz zur Gesamtheit der Publikationen weisen die referierten Beiträge eine höhere durchschnittliche Anzahl von Autoren von 4.08 auf.

2.4. Probleme

Da beim Hinzufügen neuer Einträge in die Datenbank keine Prüfung der Eingabe stattfindet, sind viele Attribute leer, oder zum Teil mit falschen oder variierenden Angaben gefüllt. Einige dieser Eingaben können bei der späteren Verarbeitung zu Problemen führen und werden deshalb in diesem Abschnitt genauer besprochen.

Die Angabe der veröffentlichenden Universität in *opus* unterscheidet sich zwischen der ausgeschriebenen Form *Technische Universität Chemnitz* (19869 Einträge) und der Abkürzung *TU Chemnitz* (547 Einträge), sowie einem Eintrag mit Schreibfehler. Da in dieser Datenbank nur Veröffentlichungen der TU Chemnitz vorhanden sind, können alle Einträge ohne entsprechende Anpassung verwendet werden. Sollten jedoch zukünftig Veröffentlichungen anderer Universitäten eingebunden werden, sollte dieses Attribut einheitlich vergeben werden.

Genau wie die Universität finden sich, um einiges vielfältiger, stark unterscheidende Namensangaben für die Autoren der Veröffentlichungen. So werden die Vornamen der Autoren einerseits ausgeschrieben, andererseits durch Initialen ersetzt. Weiterhin wurde bei der Verwendung von Initialen eine unterschiedliche Nutzung von Leerzeichen, Punkten und Kommata festgestellt. Insbesondere bei ausländischen Namen

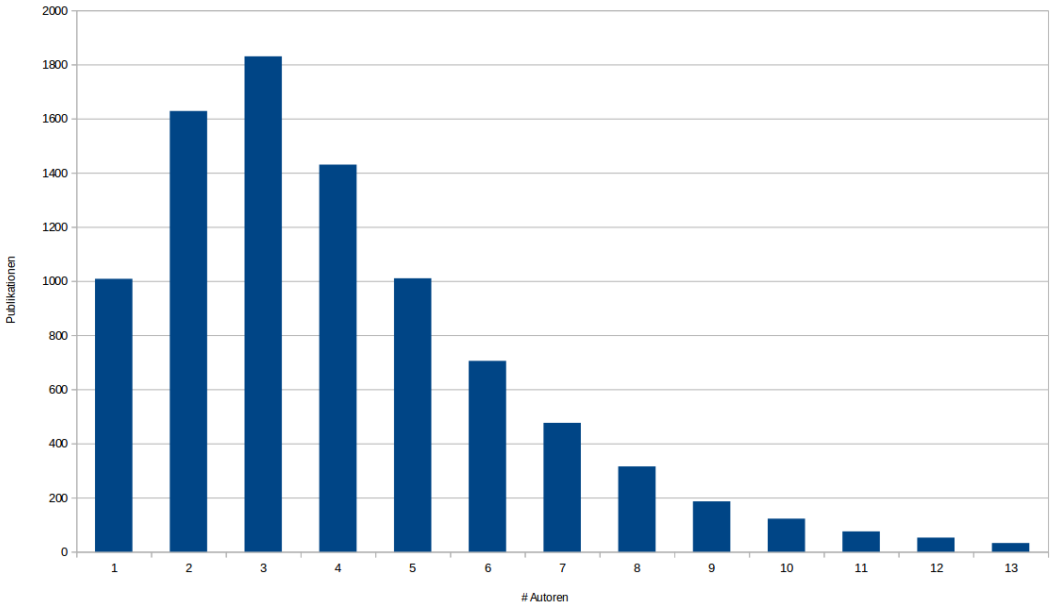


Abbildung 2.2.: Anzahl Autoren pro Publikation. Das Bar Chart zeigt die Verteilung der Anzahl von Autoren pro Publikation für referierte Arbeiten. Der Graph wurde an der x-Achse beschränkt, sodass 60 Publikationen nicht im Graphen zu sehen sind. Eine Publikation hält den Rekord mit 64 beteiligten Autoren. Wie man erkennen kann steigt der Graph bis zu einem Maximalwert von 1830 Publikationen mit 3 Autoren und sinkt anschließend logarithmisch.

fällt die unterschiedliche Nutzung von diakritischen Zeichen auf, da diese teils unterschiedlich gesetzt, oder gänzlich auf diese verzichtet wurde. In einem Einzelfall wurde der Name des Autors nicht einmal eingetragen. Der beste Ansatz zur Beseitigung der dadurch womöglich entstehenden Fehler wäre die Anpassung der Datenbankeinträge, sowie die Prüfung neuer Einträge auf Richtigkeit. Insbesondere der Vorschlag von bereits eingetragenen Namen für Neueinträge könnte Eingabefehler vermeiden und stets zur selben Schreibweise führen. Ebenfalls möglich wäre eine Relation mit Autoren, welche alle an der Universität tätigen Personen aufführt und nur diese zur Auswahl zulässt. Dies wäre aus Datenbanksicht in jedem Fall empfehlenswert, da eine n-m-Beziehung mithilfe einer Verknüpfungsrelation dargestellt werden sollte, wie dies auch bei der Relation *opus_inst* der Fall ist. Da die Daten jedoch in der gegebenen Form existieren und kein schreibender Zugriff auf diese gewährt wird, haben wir in erster Linie die gröbsten Fehler beseitigt, indem die Namen von allen Sonder-

zeichen bereinigt und anschließend gleichnamige Autoren zu einem zusammengefügt wurden. Hierbei sollte man sich dessen bewusst sein, dass trotz dieser Maßnahme einige Einträge desselben Autors nicht auf diesen zurückgeführt werden können (z.B. bei Buchstabenverdrehern) und in seltenen Fällen sogar unterschiedliche Autoren zu einem zusammengefasst würden. In Abschnitt 2.6 wird jedoch gezeigt, dass für die von uns erzeugte Darstellung die Zuordnung von Publikationen zu Kostenstellen ausreichend ist und somit nicht auf das Problem der Fehleingaben von Autorennamen eingegangen werden muss.

Einige Veröffentlichungen scheinen mehrfach in der Datenbank abgespeichert worden zu sein, während andere sehr einfache, beziehungsweise nicht aussagekräftige Titel wie *Einleitung*, oder *Studienbücher Wirtschaftsmathematik* tragen und aus diesem Grund mehrfach in der Datenbank aufzufinden sind.

Wurde ein Dokument mehrfach in die Datenbank eingeschleust, so führt dies zur Verfälschung der Daten und somit auch zu einer fehlleitenden Visualisierung. Diese würde zwar die gegebene Datenlage, jedoch nicht den realen Datenstand, korrekt widerspiegeln. Die Verwendung ausschließlich referierter Publikationen kann dabei Abhilfe schaffen.

Zwar ist das Jahr der Veröffentlichung ein Pflichtfeld, jedoch gibt es auch hier Falscheintragungen, wie beispielsweise 14 Veröffentlichung aus dem Jahr 201.

Der Umfang der Arbeiten lässt sich schwierig erfassen, da die Art der Eintragung nicht normalisiert ist. So gibt es beispielsweise die Einträge *200*, *200 S.*, *200-203*, *200, 31*, *V1-66* und *XVII, 162*. Wie man sehen kann ist auch die Bedeutung des hinterlegten Umfangs nicht immer eindeutig.

Die Internationale Standardbuchnummer (ISBN) soll normalerweise zur eindeutigen Kennzeichnung einer Arbeit dienen. In der hier gegebenen Datenbank sind diese jedoch häufig mehrfach, und im Einzelfall sogar 84 mal, an verschiedene Veröffentlichungen vergeben.

Das Feld zur Eintragung kooperierender Unternehmen wurde zum Teil zur Eintragung von Patentnummern missbraucht.

Viele Attribute könnten eine gute Grundlage zur Kategorisierung und Analyse der gespeicherten Publikationen bieten, wurden jedoch so selten verwendet, dass ihre Aussagekraft sehr niedrig ist. So wurden selbst einfach zu füllende Attribute wie die Sprache, in welcher die Veröffentlichung geschrieben wurde, nur bei 103 Einträgen angegeben.

2.5. Verwendbare Daten

Das Ziel dieser Arbeit ist es, anhand der gegebenen Publikationsdaten zu zeigen, wie gut ein Institut oder eine Fakultät im Vergleich zu anderen Instituten oder Fakultäten kooperiert. Im Folgenden wird gezeigt, wie die in 2.2 beschriebene Datenbank

verwendet wurde, um die, für die spätere Visualisierung benötigten, Informationen zu Kooperationen eines jeden Instituts zu erhalten.

Anhand der zugrundeliegenden Datenbasis wird folgend eine Kooperation dermaßen definiert, dass an der gegebenen Publikation mindestens zwei verschiedene Institute beteiligt waren. Es sei darauf hingewiesen, dass jedes andere, an einer Publikation teilnehmende, Institut jeweils als einzelner Kooperationspartner gezählt wird und somit mehrere Kooperationen pro Veröffentlichung möglich sind. Zum Beispiel werden bei einer Publikation zwischen vier verschiedenen Instituten jedem beteiligten Institut drei Kooperationen gutgeschrieben, da jedes dieser Institute bei der gegebenen Veröffentlichung drei Kooperationspartner hatte. Somit findet eine dermaßen definierte Kooperation stets zwischen zwei Instituten statt.

Die soeben definierte Kooperation lässt sich anhand der zwei beteiligten Institute weiter untergliedern in eine fakultätsinterne Kooperation, bei welcher beide kooperierenden Institute derselben Fakultät angehören, eine interfakultäre Kooperation, bei der die Kooperationspartner verschiedenen Fakultäten angehören und externen Kooperationen, bei welcher einer der beiden Kooperationspartner nicht der Universität zugehörig ist.

Um nicht nur die Anzahl der Kooperationen zwischen Instituten zu vergleichen, sondern auch das Verhältnis von Kooperationen zu Einzelarbeiten darstellen zu können, werden auch die Zahlen erfasst, bei denen ein Institut nicht kooperierte, also nur Autoren eben diesen Instituts an der Veröffentlichung beteiligt waren.

Die so erstellten Kooperationsdaten können zusätzlich nach Jahren und Dokumentarten gruppiert werden.

2.6. Extraktion der Kooperationsdaten

In diesem Abschnitt wird gezeigt, in welcher Reihenfolge welche Mengenoperationen auf der Datenbank durchgeführt werden müssen, um die in Abschnitt 2.5 festgelegten Daten zu extrahieren.

Die Zuordnung der Publikationen zu beteiligten Instituten kann auf zwei verschiedene Arten geschehen. Die offensichtliche und im Datenbankschema 2.1 gut erkennbare Variante nutzt die Einträge der Relation *opus_inst*, da diese genau für den Zweck geschaffen wurde. Dafür wird zuerst ein Equi-Join zwischen *opus_inst* und *institute_de* auf das Attribut *inst_nr* durchgeführt, wodurch jedem Eintrag die passende Kostenstelle zugeordnet wird. Anschließend folgt ein weiterer Equi-Join der entstandenen Relation mit sich selbst auf das Attribut *source_opus*, wodurch man alle Kombinationen von Instituten erhält, welche an derselben Veröffentlichung mitgewirkt haben. Je nachdem, welche Art von Kooperation herausgefunden werden soll, kann nun eine weitere Selektion anhand der beiden vermerkten Fakultätsangaben erfolgen. Möchte man die Kooperationen später an einer Zeitleiste ausrichten, so wird ein weiterer

Equi-Join mit *opus* auf das Attribut *source_opus* benötigt, wodurch jedem Eintrag das Veröffentlichungsjahr und die Dokumentart zugeordnet wird. Je nachdem, ob die jeweiligen Kooperationspartner von Bedeutung sind, oder nur der Absolutwert der gewünschten Kooperationen pro Institut in Erfahrung gebracht werden soll, erfolgt eine Gruppierung inklusive einer Zählung nach beiden Kostenstellen, oder nur nach der ersten Kostenstelle.

Da externe Mitarbeiter keinem Institut der Universität zugeordnet werden können, sind externe Zuarbeiten nur in der Relation *opus_autor* zu finden, in welcher als Kostenstelle *extern* angegeben wird. Um die Kooperationsdaten zu extrahieren muss die Relation zuerst auf die Beziehung von Instituten zu Veröffentlichungen reduziert werden, anstatt Beziehungen von Autoren zu Veröffentlichungen darzustellen. Diese Reduktion kann durch die Gruppierung nach den beiden Attributen *source_opus* und *kst* erfolgen. Würde man an dieser Stelle alle externen Beziehungen entfernen, so erhielte man theoretisch dieselben Einträge wie in *opus_inst*; praktisch findet sich jedoch eine Diskrepanz von 20 Einträge zugunsten der hier beschriebenen Methode. Woher diese Diskrepanz stammt konnte im Laufe dieser Arbeit nicht geklärt werden. Anschließend wird ein Equi-Join der entstandenen Relation mit sich selbst auf das Attribut *source_opus* durchgeführt, wodurch man alle Kombinationen von Instituten erhält, welche an derselben Veröffentlichung mitgewirkt haben. Um die gewünschten externen Mitarbeiten darzustellen, wird eine Selektion durchgeführt, welche prüft, ob die zweite angegebene Kostenstelle extern ist. Möchte man zu jedem Eintrag das Veröffentlichungsjahr und die Dokumentart kennen, wird ein weiterer Equi-Join mit *opus* auf das Attribut *source_opus* benötigt. Um den Absolutwert der externen Kooperationen pro Institut zu erhalten, erfolgt eine Gruppierung inklusive einer Zählung nach der Kostenstelle.

Um alle Arbeiten aufzulisten, welche von einem einzigen Institut veröffentlicht wurden und somit nicht Teil einer Kooperation sind, werden die Einträge aus *opus_inst* selektiert, deren *source_opus*-Attribut nicht in der Liste der oben erstellten Kooperationsrelation zu finden ist.

2.7. Vollständige oder partielle Zählung

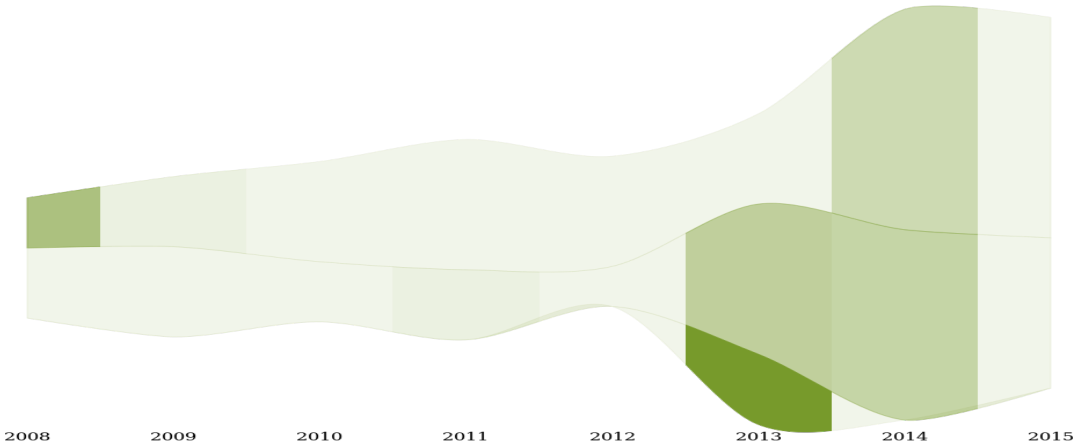
Auf Basis der Datenanalyse in Kapitel 2.5 stellte sich heraus, dass die Beziehungen des Netzwerkes nicht anhand von Zitierungen, sondern ausschließlich anhand der Mehrautorenschaft hergestellt werden können. Da für Autoren keine Relation in der Datenbank existiert und somit bereits einfache Schreibfehler zu Problemen führen (siehe Abschnitt 2.4) und weiterhin ein Autor im Laufe seiner Arbeit an verschiedenen Instituten publizieren kann, werden die Autoren einer Veröffentlichung in unserer Anwendung auf ihre an der Arbeit beteiligten Institute heruntergebrochen. Da jedes Institut auch einer Fakultät zugeordnet ist, können sowohl die Beziehungen zwischen

Instituten, als auch zwischen Fakultäten dargestellt werden.

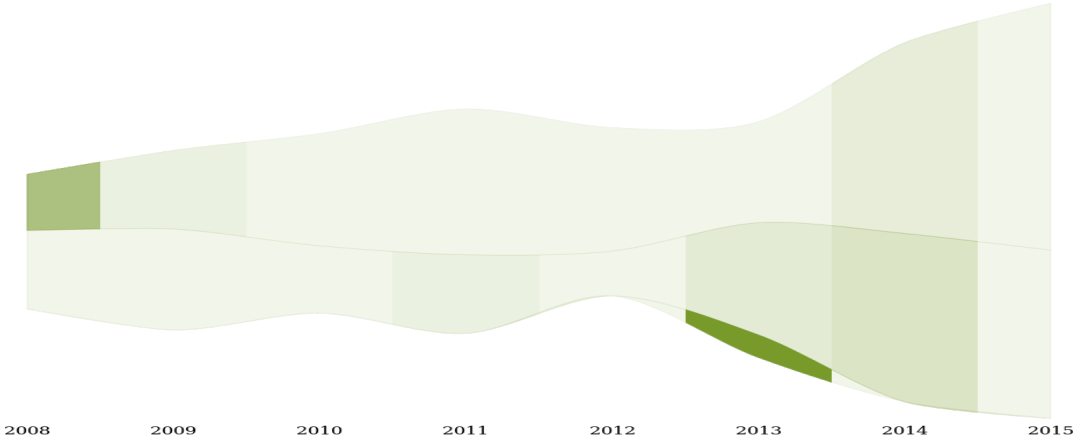
In der Arbeit [PRWvE16] wird darauf hingewiesen, dass die Frage, ob Kooperationen vollständig oder partiell gezählt werden, eine entscheidende Auswirkung auf die Darstellung und die Aufnahme der Informationen hat. Partielles Zählen hilft, den Einfluss von Publikationen mit vielen Koautoren geringer zu halten. Hat eine Publikation beispielsweise Autoren von 20 Instituten, so werden jedem dieser Institute 19 Kooperationen allein für diese Veröffentlichung angerechnet, womit in der Visualisierung jedes dieser Institute mehr ins Auge fällt als ein Institut, welches 10 Publikationen mit jeweils einem anderen Institut veröffentlichte. Dieser Fakt fällt umso mehr ins Gewicht, wenn die Kooperationen nach ihrem Veröffentlichungsjahr unterteilt werden, da dann alle 19 Kooperationen auf dasselbe Jahr fallen, während die 10 Kooperationen über mehrere Jahre verteilt sein können. Der Nachteil der partiellen Zählung ist, dass die Kooperation vieler Institute an derselben Veröffentlichung komplett aus dem Fokus gerückt wird, weil sie nun genau denselben Wert hat wie die Kooperation zwischen zwei Instituten. Der Unterschied zwischen vollständiger und partieller Zählung wird in Abbildung 2.3 dargestellt. Die unterste Schicht in der Darstellung repräsentiert das Institut *Interdisziplinäres Kompetenzzentrum*, welches laut [int11] den Anspruch der gemeinsamen Arbeit von Forschern unterschiedlicher Fachrichtungen hat, deshalb in Anbetracht der Kooperationen von Instituten eine entscheidende Rolle spielt und einen entsprechenden Anteil an der Visualisierung haben sollte. Das Institut war im Jahr 2013 nur an einer Publikation beteiligt, welche jedoch in Zusammenarbeit von acht Instituten entstand. Je nach Zählung wird dem Institut entweder ein Wert von eins, oder von sieben zugesprochen und die Schicht entsprechend dick dargestellt, wodurch die Schicht bei partieller Zählung in der Darstellung untergeht. Die Verwendung eines Gradienten, welcher in Kapitel 4.9 eingeführt wird, zur Hervorhebung von Kooperationen innerhalb und zwischen Fakultäten kann die Schicht dennoch ins Auge des Betrachters rücken, da die in Kooperation entstandenen Arbeiten in beiden Fällen einhundert Prozent der Publikationen dieses Instituts ausmachen und die Farbe somit ohne Transparenz erscheint. An dieser Stelle wird deutlich, dass auch bei größeren Instituten Probleme bei partieller Zählung auftreten, da die Anzahl an Kooperationen geringer wird, aber die der Publikationen innerhalb des Instituts und mit externen Mitarbeitern unverändert bleiben. Zwar spiegelt dies das Verhältnis von tatsächlich veröffentlichten Arbeiten besser wieder als die Zählung jeden Instituts, mit dem kooperiert wurde, andererseits sollen in unserer Anwendung stärker kooperierende Institute in den Vordergrund geschoben werden. Aus diesem Grund wurde sich in dieser Anwendung für die vollständige Zählung entschieden.

Normalerweise führt die partielle Zählung dazu, dass die gezählten Kooperationen mit der Anzahl der Veröffentlichungen eines Instituts übereinstimmt, weil die gewichteten Kooperationen sich zu eins aufaddieren und somit auch die absoluten Publikationszahlen untersucht werden können. Da die Kooperationen in unserem Fall jedoch in drei Kategorien unterteilt werden, um zu sehen, mit wem kooperiert wurde, kann

eine Veröffentlichung pro Kategorie ein mal gezählt werden und weicht damit von den absoluten Publikationszahlen ab. Dem könnte man entgegenwirken, indem zusätzlich zur partiellen Zählung der Kooperationen die Anteile der Kategorien eingerechnet werden.



(a) Vollständige Zählung



(b) Partielle Zählung

Abbildung 2.3.: Vollständige oder partielle Zählung. Die Darstellungen (a) und (b) zeigen den Streamgraph für drei ausgewählte Institute der Informatik in den Jahren 2008 bis 2015. Der Gradient zeigt prozentual das Verhältnis des Wertes von intra- und interfakultären Kooperationen zum Gesamtwert der jeweiligen Schicht pro Jahr. In (a) wurden die Kooperationen derart berechnet, dass einem Institut für jede Publikation maximal eine intrafakultäre und eine interfakultäre Kooperation angerechnet werden kann, während in (b) jede Publikation einen Wert von $n-1$ erreicht, wobei n die Anzahl an der Publikation beteiligter Institute ist und der Wert, je nach Zugehörigkeit der kooperierenden Institute, auf die beiden Kooperationsarten aufgeteilt wird.

3. Visualisierung

3.1. Einleitung

Die Visualisierung von Kooperationen zwischen Instituten verschiedener Fakultäten geschieht in unserem Fall anhand der Mehrautorenschaft von Publikationen. In diesem Kapitel werden Darstellungen vorgestellt, welche für ähnliche Anwendungsfälle, wie Zitierungsnetzwerke oder Soziale Netzwerke, entwickelt wurden. Neben der Netzwerkanalyse, welche die Beziehungen zwischen einzelnen Akteuren herausarbeitet, wird Wert darauf gelegt, ob die gegebene Darstellung die in Kapitel 1.2 gesteckten Ziele erfüllt und ob sie in der Lage ist, oder dermaßen erweitert werden kann, dem Anwender Einblick in die zeitliche Entwicklung der Akteure und ihrer Beziehungen zu geben. Schließlich wird eine weitere Visualisierung vorgestellt, welche nicht zur Netzwerkanalyse, aber spezifisch zur Darstellung zeitlich bedingter Daten geeignet ist und welche letztendlich die Basis für die spätere Implementierung darstellt.

3.2. Aufbau eines Netzwerkes

Um eine Netzwerkanalyse auf Publikationsdaten durchführen zu können, muss erst entschieden werden, auf welche Art und Weise die gegebenen Daten in ein Netzwerk umgeformt werden.

Prinzipiell kann ein Netzwerk mithilfe eines ungerichteten, gewichteten Graphen $G = (V, E)$ modelliert werden, welcher aus Knoten V und Kanten E besteht, wobei jede Kante zwei Knoten miteinander verbindet $E = (v, u); v, u \in V$ und jeder Kante ein nicht-negatives Gewicht w zugeordnet werden kann.

Je nach Auslegung kann ein Knoten eine Fakultät, ein Institut, ein Autor, oder eine Publikation symbolisieren, wobei es ebenfalls möglich ist diese zu kombinieren, indem verschiedene Knoten je eine dieser Entitäten repräsentieren. Je nach Datenbasis könnten diese auch Journals, oder verwendete Schlüsselworte darstellen [vEW14b]. Eine Kante stellt jeweils eine Beziehung zwischen zwei Knoten her, wobei die Beziehung vom Typ der Knoten abhängig ist. Wenn die Kooperationen wie in unserem Fall auf der Basis von Mehrautorenschaft berechnet werden, so kann eine Kante zwischen zwei gleich gearteten Knoten (zwei Fakultäten, zwei Instituten, oder zwei Autoren) darauf hinweisen, dass zwischen diesen eine Publikation in Zusammenarbeit entstanden ist. Würde die Darstellung stattdessen auf Zitierungen basieren, würden diese eher als gerichtete Kanten zwischen Publikationen dargestellt werden. Auch Beziehungen auf

Basis derselben verwendeten Schlüsselworte wäre möglich. Zusätzlich zu Kooperationen können hierarchische Zugehörigkeiten im Graphen festgehalten werden, indem ein Institut einer Fakultät, ein Autor einer oder mehreren Institutionen, oder eine Publikation einem oder mehreren Autoren zugeordnet wird. Genauso könnten hierarchische Stufen übersprungen werden, indem beispielsweise eine Publikation direkt mit den beteiligten Fakultäten verbunden wird. Das Gewicht einer Kante bestimmt dabei, wie stark die Verbindung zwischen den beiden Knoten ist. Die Stärke kann unter anderem ausdrücken, dass ein Autor besonders viele Publikationen bei einem Institut veröffentlicht hat, oder zwei Autoren besonders häufig zusammengearbeitet haben.

Der Aufbau des Netzwerkes verändert schließlich die Art, wie verschiedene Algorithmen zur Darstellung des Netzwerkes agieren und welche Informationen folglich aus der Visualisierung gelesen werden können.

3.3. Graph

Die Visualisierung von Graphen wurde bereits häufig zur Darstellung von Sozialen Netzwerken [ABA03][HB05] und Publikationsdaten anhand von Zitierungen [BP11] und Mehrautorenschaft [New04] genutzt. In der Arbeit von Linton Freeman [Fre00] wird sowohl gezeigt, wie sich die Graphen entwickelt haben, als auch, inwiefern die Position, Farbgebungen, Form und Größe von Knoten genutzt werden können, um verschiedene Informationen darzustellen. Dabei unterscheiden sich insbesondere die Algorithmen zur Positionsbestimmung der Knoten, der Darstellung von Kanten und der Anzeige von Clustern oder Gruppenzugehörigkeiten innerhalb des Netzwerks.

3.3.1. Kräftewirkungen im Graphen

Die Anordnung der Knoten, und damit auch der Kanten, entscheidet über das allgemeine Erscheinungsbild des Graphen und somit über die durch den Nutzer lesbaren Informationen und kann nach verschiedenen Kriterien erfolgen. Einerseits können Ästhetische Aspekte, wie die optimale Länge von Kanten, oder die Minimierung der Zahl an Überkreuzungen, Einfluss auf die Algorithmenwahl haben [Dwy09]. Andererseits kann die Distanz zwischen Knoten genutzt werden, um deren Zusammengehörigkeit zueinander darzustellen, wobei kleinere Distanzen eine höhere Zahl an Zusammenarbeiten darstellen, oder die Position kann unabhängig von Zusammenarbeiten die Zugehörigkeit von Knoten zu bestimmten Gruppen oder Hierarchien hervorheben (zum Beispiel wenn alle Institute derselben Fakultät nah nebeneinander liegen).

Für die Positionierung von Knoten in beliebigen Graphen hat sich die Analogie zur Berechnung von Kräfteverhältnissen zwischen Teilchen durchgesetzt. [EAD84][FR91] Dabei werden sowohl, auf dem Coulombschen Gesetz basierende, abstoßende Kräfte zwischen allen Teilchen, als auch, auf dem Hook'schen Gesetz basierende, anziehende

Kräfte entlang der Kanten des Graphen simuliert. [Kob04] Dabei können verschiedenste Algorithmen zur Verarbeitung der im Graphen vorhandenen Informationen von Knoten- und Kantengewichten verwendet werden, welche nicht den realistischen physikalischen Gesetzen entsprechen müssen, sondern die Darstellung bezüglich zuvor festgelegter Kriterien optimieren.

Viele weitere Ansätze zur Berechnung der wirkenden Kräfte, zum Beispiel anhand von abstoßenden Kräften zwischen Kantenmittelpunkten [CP96], zwischen Kanten und Knoten [Ber99], oder unter Einbeziehung der Knotengrade (der Grad eines Knotens entspricht der Anzahl seiner mit ihm verbundenen Kanten) [FLM94], existieren [Noa07].

Da Kräfte über die Zeit wirken, werden diese aus einem festgelegten Grundzustand berechnet und alle Teilchen entsprechend der auf sie wirkenden Kräfte verschoben. Dies wird in so vielen Zeitschritten wiederholt, bis sich das System stabilisiert hat. Je nach verwendeten Algorithmen können sehr viele Iterationen von Nöten sein, um einen stabilen Zustand zu erreichen, weshalb in [FR91] eine *Temperatur* eingeführt wird, welche bei eins beginnt und in jedem Zeitschritt verringert wird, bis sie schließlich null erreicht. Die wirkenden Kräfte werden in jedem Zeitschritt mit der aktuellen Temperatur multipliziert, sodass die Kräfte in jedem Schritt weniger stark wirken und der Algorithmus spätestens terminiert, wenn die Temperatur null erreicht.

3.3.2. Aufwand

Da jedes Teilchen eines Systems jedes andere Teilchen beeinflussen kann, und die neue Position für jedes Teilchen berechnet werden muss, beträgt die Zahl der Berechnungen

$$\frac{1}{2}N(N-1) = O(N^2) \tag{3.1}$$

pro Zeitschritt.

Die Komplexität kann unter anderem mithilfe des Algorithmus von Barnes & Hut [BH86] auf $O(NlogN)$ reduziert werden, indem der vom Graphen maximal eingenommene Platz, in Abhängigkeit der Verteilung der Partikel, durch einen Quadtree (in 2D) hierarchisch unterteilt wird und die Kräfte der in einer Zelle befindlichen Teilchen rekursiv zusammengefasst werden. Somit muss nicht mehr jedes Teilchen mit jedem anderen verglichen werden, sondern die wirkenden Kräfte werden auf die Wechselwirkung des Teilchens mit ganzen Zellen reduziert, solange diese ein gewisses *Multipol-Akzeptanz Kriterium* (engl.: MAC; Multipole-Acceptance-Criterion) [gal] erfüllen, um größere Approximationsfehler zu vermeiden.

Wenn die Knotenzahl sehr groß ist, kann die Komplexität unter gewissen Voraussetzungen weiter gedrückt werden. [GR87]

3.3.3. Implementierung

Da zeitgleich zu der hier vorliegenden Arbeit eine Netzwerkanalyse anhand der Publikationsdaten der TU Chemnitz durchgeführt wurde [AB16], werden im Folgenden ausschließlich die Algorithmen betrachtet, welche in der dazu verwendeten Software Gephi [BHJ+09] zur Bildung des Graphen führten.

Der verwendete Algorithmus ForceAtlas2 [JVHB14] berechnet die anziehende Kraft F_a zwischen zwei durch eine Kante verbundene Knoten n_1 und n_2 linear anhand ihres Abstandes d

$$F_a(n_1, n_2) = w(e)d(n_1, n_2) = w(e)|r_2 - r_1| \tag{3.2}$$

, wobei r der Ortsvektor des entsprechenden Knotens und w(e) das Gewicht der Kante zwischen den beiden Knoten ist. Aus der Formel folgt, dass durch Kanten verbundene Knoten umso mehr voneinander angezogen werden, desto weiter sie voneinander entfernt sind, was der Analogie zu einer Sprungfeder entspricht.

Die abstoßende Kraft F_r zwischen zwei Knoten n_1 und n_2 wird anhand der Formel

$$F_r(n_1, n_2) = k_r \frac{(deg(n_1) + 1)(deg(n_2) + 1)}{d(n_1, n_2)} \tag{3.3}$$

berechnet, wobei *deg* den Grad des Knotens beschreibt und k ein einstellbarer Koeffizient ist. Aus dieser Berechnung folgt, dass Knoten sich umso mehr abstoßen, desto näher sie sich sind. Weiterhin erfahren wenig vernetze Knoten eine verhältnismäßig geringe Abstoßung, während zwei stark vernetzte Knoten sehr weit voneinander entfernt werden. Dies führt unter anderem zu einer besseren Verteilung von Gruppen innerhalb des Netzwerks.

Anstatt wie in Abschnitt 3.3.1 beschrieben eine allgemeine Temperatur zur Abkühlung der Bewegungen im Graphen zu verwenden, wird lokal in jedem Zeitschritt geschaut, wie stark sich die auf den Knoten wirkenden Kräfte von diesem Zeitschritt im Vergleich zum vorigen verändert haben und anhand dieses Wertes wird die Bewegung eines jeden Knotens umso mehr entschleunigt, desto stärker er schwingt.

Um die Knoten, insbesondere jene ohne Kanten, innerhalb der Darstellung zu halten wird eine *Gravitation* eingeführt, welche alle Knoten ins Zentrum der Darstellung zieht. Die auf einen Knoten wirkende Gravitation wird in dieser Implementierung vom Knotengrad abhängig gemacht, um stärker vernetzte Knoten in den Fokus der Darstellung zu ziehen.

$$F_g(n) = k_g(deg(n) + 1) \tag{3.4}$$

Der Algorithmus wird weiterhin mithilfe der in Abschnitt 3.3.2 beschriebenen Barnes & Hut Approximation beschleunigt.

In der Arbeit [AB16] wurde das Publikationsnetzwerk derart erstellt, dass jede Publikation durch einen kleinen goldenen Knoten und jeder Autor durch einen mittelgroßen Knoten in der Farbe der Fakultät, in der er am meisten veröffentlichte

repräsentiert wird. Fakultäten werden in ihre Institute unterteilt und jedes Institut durch einen großen Knoten in der Farbe entsprechend des Cooperate Design [coo14] dargestellt. Fakultäten ohne Institute werden nur durch einen Knoten repräsentiert. Eine Kante existiert sowohl zwischen einer Publikation und ihren Autoren, als auch zwischen einem Autor und den Instituten, beziehungsweise den Fakultäten, an denen er veröffentlichte. Die beschriebene Struktur wird in Abbildung 3.1 gezeigt.

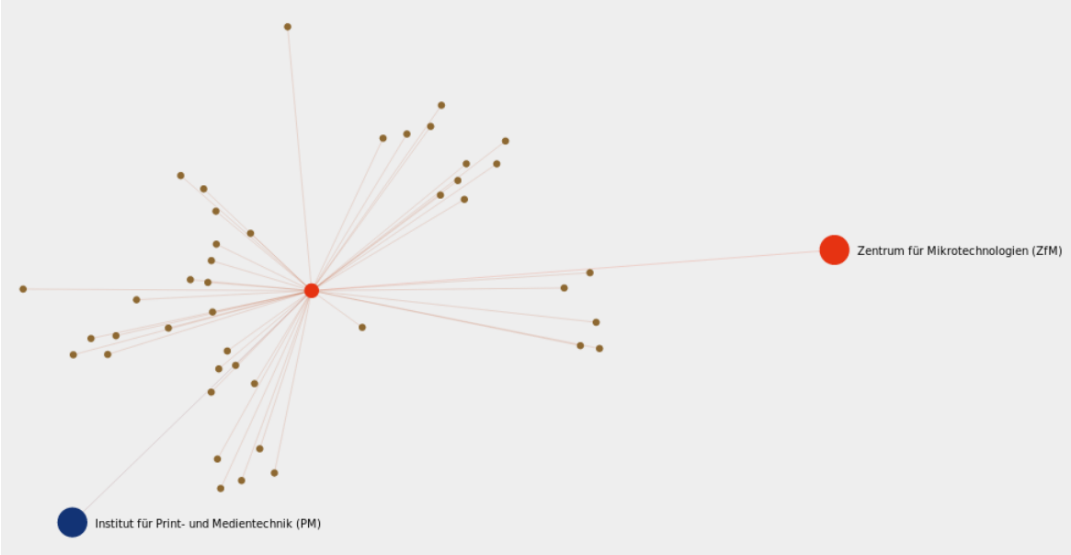


Abbildung 3.1.: Struktur des Graphen. Mittelgroße, farbige Knoten repräsentieren einen Autor, welcher mit den Instituten, an denen er veröffentlichte (große farbige Knoten), durch eine Kante verbunden wird. Die kleinen, goldenen Punkte repräsentieren alle Publikationen, an denen der Autor beteiligt war. Die Farben entsprechen denen des Cooperate Design [coo14]

3.3.4. Auswertung

Betrachtet man den Graphen in Abbildung 3.2, so lassen sich sowohl allgemeine Aussagen über das Netzwerk treffen, als auch spezifische Fragen zu Beziehungen zwischen den Instituten beantworten.

Mithilfe des Graphen kann grob die Größe einer jeden Fakultät abgeschätzt werden, indem die Menge an gleichfarbigen Knoten verglichen werden. So lässt sich erkennen, dass die Fakultät für Maschinenbau (dunkelblau) vermutlich die größte Fakultät der Universität ist, gefolgt von der Fakultät für Elektrotechnik und und Informations-

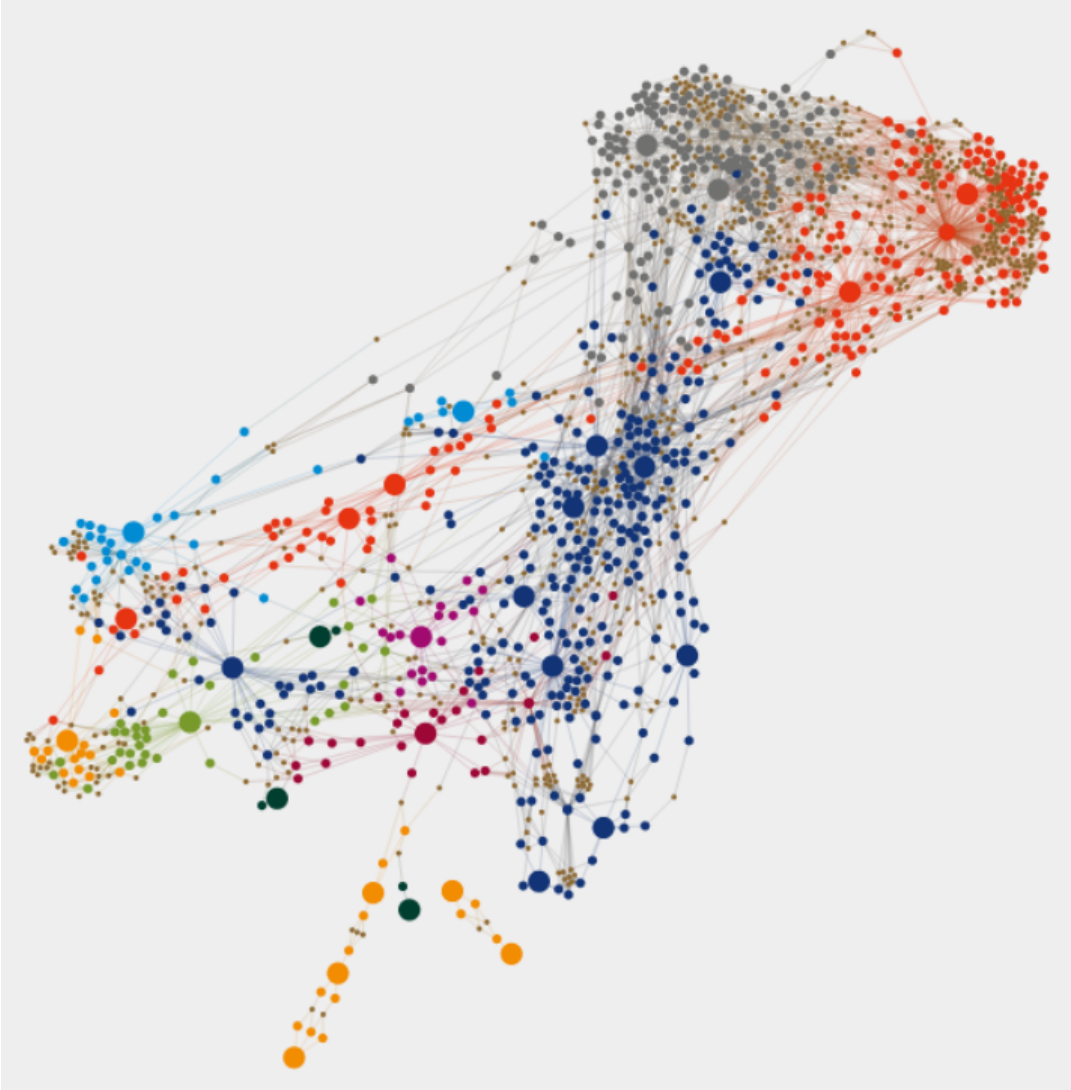


Abbildung 3.2.

technik (leuchtorange) und der Fakultät für Naturwissenschaften (grau), welche beide ungefähr gleich groß zu sein scheinen. Da die Menge an farbigen Knoten von der Anzahl der veröffentlichenden Autoren abhängt, kann an dieser Stelle ein direkter Bezug zur Anzahl der Mitarbeiter gezogen werden. Wie man jedoch an der Philosophischen Fakultät (dahliengelb) und der Fakultät für Human- und Sozialwissenschaften (hellblau) erkennt, ist die Abschätzung der Größe von Fakultäten umso schwieriger, desto

mehr deren Knoten auseinandergezogen werden. Zusätzlich zur allgemeinen Größe einer Fakultät ist auch direkt ablesbar, in wie viele Institute diese untergliedert wurde, da diese jeweils durch einen großen Knoten dargestellt werden.

Aufgrund der in 3.3.1 beschriebenen, wirkenden Kräfte, werden Publikationen zwischen den aktuellen Positionen der Autoren angeordnet, welche an dieser Veröffentlichung beteiligt waren. Weiterhin werden Autoren zwischen ihren Publikationen und den Instituten, an denen sie veröffentlichten, positioniert. Die Institute stellen stets den Mittelpunkt der bei ihnen veröffentlichenden Autoren dar. Da in dieser Implementierung alle Kantengewichte gleich groß sind, wird die Position eines Knotens aus dem Mittel der Positionen der mit ihm verbundenen Knoten gebildet. Die Abbildung 3.3 zentriert den Graphen auf eine Publikation, welche die einfachste solcher Beziehungen zeigt. Die Publikation (goldener Punkt in der Mitte der Darstellung) wurde von genau zwei Autoren verschiedener Fakultäten veröffentlicht, welche jeweils nur diese eine Publikation veröffentlichten. Dadurch liegt der Knoten der Publikation sowohl mittig zwischen den Autorknoten, als auch mittig zwischen den beiden Institutsknoten. Wären stattdessen drei Autoren beteiligt gewesen, von denen zwei demselben Institut angehören, so wäre die Publikation im Verhältnis eins zu zwei in Richtung der beiden Autoren desselben Instituts gerutscht. Weiterhin wären alle Knoten näher zu einem der beiden Institute gerückt, wenn einer der beiden Autoren an einer weiteren Publikation beteiligt wäre, welche nur an jenem Institut veröffentlicht wurde.

Die Lage der Knoten drückt schließlich die Beziehung der Institute zueinander in Hinsicht auf ihre Veröffentlichungen aus. Die Kooperationen lassen sich wie bereits in 2.5 erwähnt in drei Gruppen gliedern: innerhalb des eigenen Instituts, intrafakultär und interfakultär. Wenn viele Publikationen innerhalb des eigenen Instituts und der eigenen Fakultät entstanden, bilden die Autorenknoten einfarbige Cluster. Erst durch interfakultäre Veröffentlichungen werden verschiedenfarbige Knoten voneinander angezogen. Aus der Abbildung 3.2 lässt sich zum Beispiel ablesen, dass die Fakultät für Naturwissenschaften (grau) und die Fakultät für ET und IT (leuchtorange) deutliche Cluster bilden und somit viel innerhalb der eigenen Fakultät publizieren. Da sich diese Cluster jedoch sehr nah sind, scheinen auch viele Arbeiten in Kooperation miteinander entstanden zu sein. An diesen Publikationen scheint weiterhin ein Institut der Fakultät für Maschinenbau (Institut für Print- und Medientechnik) beteiligt gewesen zu sein.

Da weniger vernetze Knoten nach der Gleichung 3.4 eher an den Rand der Visualisierung gedrängt werden, bedeutet dies, dass die Knoten mit den meisten interfakultären Kooperationen in der Mitte des Graphen liegen. So scheint die Fakultät für Mathematik (violett) sehr gut interdisziplinär vernetzt zu sein, während die Philosophische Fakultät kaum kooperiert (bis auf das Institut für Medienforschung, welches sehr nah an der Fakultät für Informatik (grün) liegt, jedoch vom eigentlich Netz der Philosophischen Fakultät abgetrennt ist).

Wenn Autoren einer Fakultät den Großteil ihrer Arbeiten mit einzelnen anderen

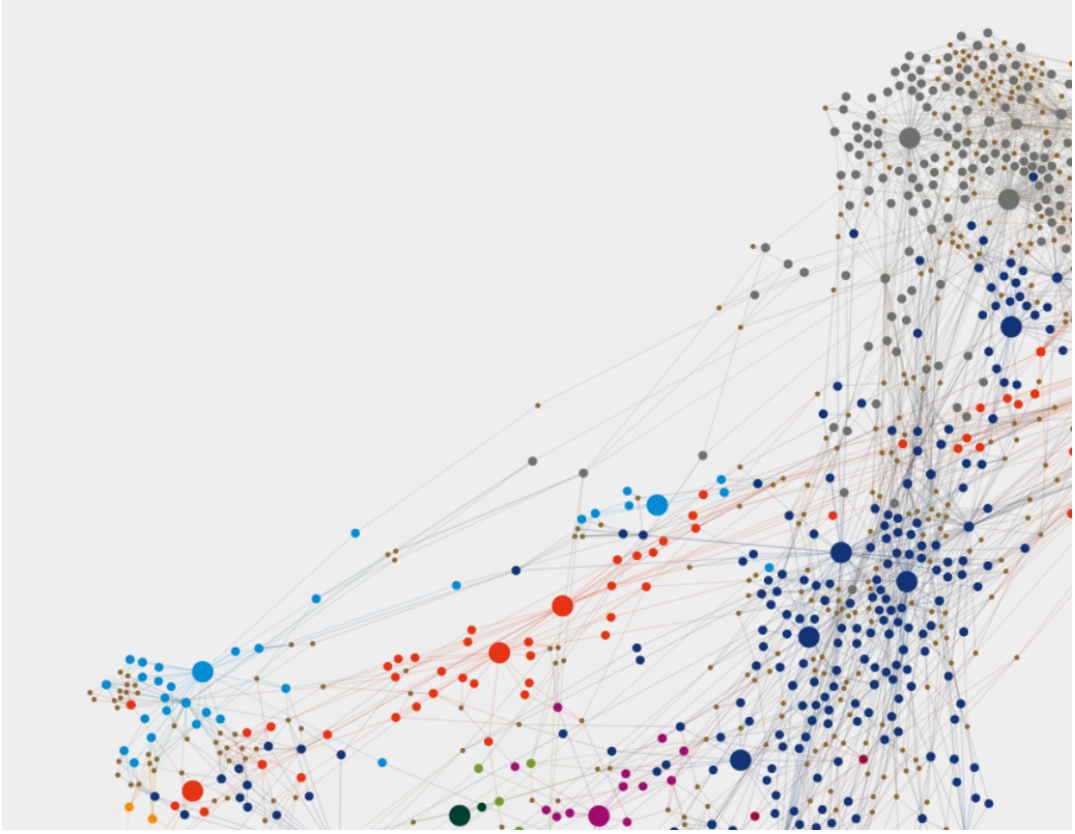


Abbildung 3.3.: Einfache Demonstration der Kräftewirkung. Der Graph wurde auf eine Publikation (goldener Knoten in der Mitte der Darstellung) zentriert, welche je eine Kante zu ihren beiden Autoren (mittelgroßer hellblauer und grauer Knoten) besitzt, welche wiederum durch eine Kante mit ihren Instituten verbunden wurden. Aufgrund der symmetrischen Kräftewirkung liegt der Knoten der Publikation mittig zwischen den Autoren und ihren Instituten.

Fakultäten kooperieren, werden deren Knoten zum jeweiligen Kooperationspartner gezogen und somit stark über den kompletten Graphen verteilt. Dadurch sind solche Fakultäten deutlich unauffälliger als jene, welche nur innerhalb der eigenen Fakultät kooperieren, da diese zu großen Clustern vereinigt werden. Dieser Effekt sorgt dafür, dass interdisziplinäre Fakultäten weniger auffällig sind, als andere, was entgegen den ursprünglichen Zielsetzungen dieser Arbeit steht, Beziehungen zwischen verschiedensten Instituten hervorzuheben.

Der Graph gibt einen guten Überblick über die Verteilung und Zusammenarbeit der Autoren verschiedener Institute. Aufgrund seiner visuellen Beschaffenheit fallen jedoch hauptsächlich gebildete Cluster auf, da diese eine signifikante Fläche gleichmäßig einfärben, als auch einzelne Knoten, welche sich innerhalb der Cluster anderer Fakultäten befinden, da sie sich farblich von der Masse abheben. Autoren und Institute, welche an mehreren interdisziplinären Arbeiten verschiedener Fakultäten beteiligt waren, liegen in der Mitte des Graphen zwischen den großen Clustern und werden direkt neben andere, ähnlich kooperative, Knoten gelegt. Die Ansammlung verschiedenfarbiger Knoten, welche womöglich nichts miteinander zu tun haben, sondern aufgrund ihrer ähnlichen Kooperationen ähnlich positioniert wurden, ist nur schwer leserlich und rückt durch ihre Undurchsichtigkeit die interdisziplinären Knoten aus dem Fokus heraus. Würde dem Graphen ein Knoten für externe Mitarbeiter hinzugefügt werden, würden die herausgearbeitete Struktur und Gruppierungen verloren gehen, weil nahezu jedes Institut mit externen Mitarbeitern kooperiert.

3.3.5. Erweiterungen

Nach der Gleichung 3.3 stoßen Knoten andere Knoten umso mehr ab, desto mehr Kanten mit ihnen verknüpft sind. Mithilfe dieses Wissens lassen sich Autoren finden, welche an besonders vielen Publikationen beteiligt waren, da um sie herum mehr freier Platz ist (siehe Abbildung 3.4). Jedoch liegen auch viele Autorenknoten zufällig an relativ freien Stellen, mit dem Unterschied, dass von ihnen weniger sichtbare Kanten weggehen. Um solche Autoren mehr in den Vordergrund zu rücken, könnten deren Knoten mit der Anzahl an beteiligten Publikationen skaliert werden. Dies hätte weiterhin den Vorteil, dass einzelne Autorknoten innerhalb des Clusters einer anderen Fakultät (in Abbildung 3.4 liegen einige vereinzelte dunkelblaue Knoten innerhalb des grauen Clusters) danach unterschieden werden könnten, ob sie dort positioniert wurden, weil sie an so vielen Publikationen mit Autoren dieser Fakultät kooperierten, oder weil die wenigen Publikationen, an denen sie mitarbeiteten, so viele Koautoren aus jener Fakultät hatten.

Die Kanten des Graphen dienen der Berechnung von Kräften zwischen den Knoten und damit deren Positionierung. Die visuelle Darstellung der Kanten in dieser Form gibt dem Nutzer jedoch kaum zusätzliche Informationen, da das Netzwerk bereits durch die Verteilung der Knoten gut strukturiert wurde. Nur in Einzelfällen helfen die Kanten dabei, die Beziehungen den richtigen Knoten zuzuordnen, jedoch führt die Überschneidung und Menge der Kanten zu einem visuellen Durcheinander (*Overplotting*). Um die Menge an Kanten zu reduzieren und ihnen informellen Mehrwert zu verleihen, indem sie speziell die Beziehungen zwischen verschiedenen Gruppen darstellen und den Wert der Beziehung visuell erfassbar machen, können Algorithmen zum *Edge Bundling* verwendet werden, welche mehrere ähnliche Kanten zu einer größeren zusammenfassen.

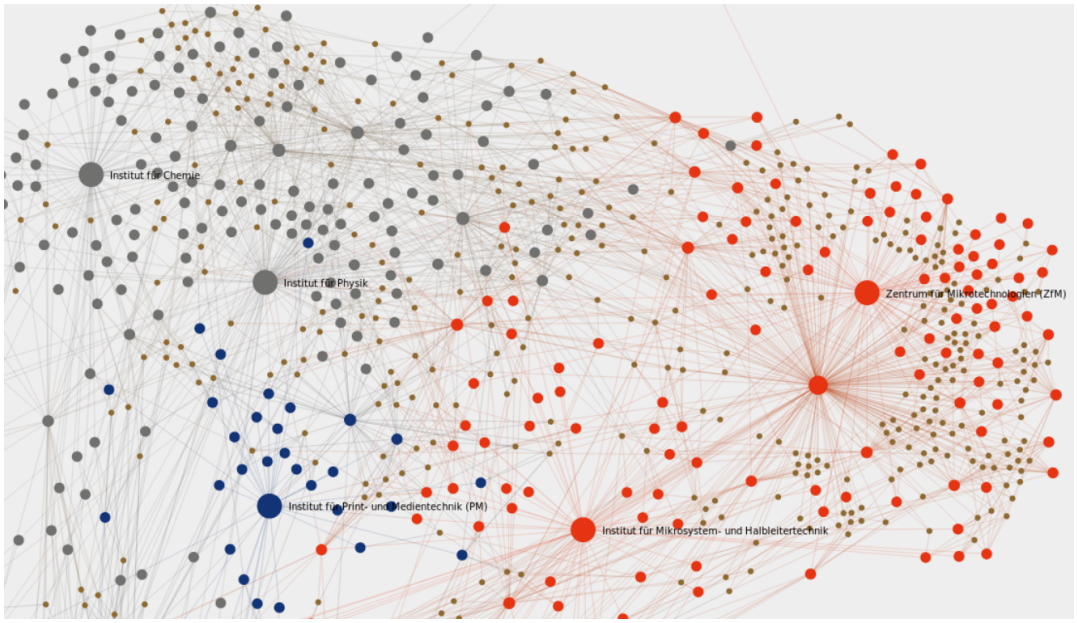


Abbildung 3.4.: Herangezoomter Graph. Knoten, welche mit vielen Kanten verbunden sind, stoßen andere Knoten stärker ab und können dadurch visuell auffallen. Vereinzelte blaue Knoten liegen im grauen Cluster, wobei nicht klar ist, ob sie viele Publikationen mit dieser Fakultät veröffentlichten, oder nur viele Autoren dieser Fakultät an den Publikationen beteiligt waren. Beide Aspekte könnten davon profitieren, wenn die Knoten mit der Anzahl ihrer Publikationen skaliert würden.

3.4. Graphen mit zeitlichem Bezug

Entsprechend der in 1.2 definierten Zielstellungen soll die Entwicklung des Netzwerkes, einzelner Institute und ihrer Beziehungen über die Zeit aus der Visualisierung extrahierbar sein. Da jeder Publikation ein Veröffentlichungsjahr zugeordnet wird, sind die dafür benötigten Informationen in der Datenbasis enthalten.

Um die zeitliche Entwicklung in einem Graphen darzustellen, wurden bereits verschiedene Methoden veröffentlicht und werden im Folgenden vorgestellt und auf ihre Verwendbarkeit hin untersucht.

3.4.1. Zeitebenen

Die vermutlich einfachste Idee ist es, die Publikationen nach ihrem Erscheinungsjahr zu gruppieren und für jedes Jahr den Graphen zu berechnen. Die einzelnen zweidimensionalen Bilder können dann in der dritten Dimension an einer Zeitleiste ausgerichtet, übereinander gelegt, oder parallel betrachtet werden. [ITK10] Wie man in Abbildung 3.5 erkennen kann, treten in (a) je nach Blickwinkel Überlagerungen der Ebenen auf und es kommt zu perspektivischer Verzerrung, welche die Lesbarkeit der Informationen beeinflusst. Desto breiter ein Graph ist, desto größer sind die beiden Störfaktoren. Werden die Ebenen wie in (b) übereinander gelegt, so wirkt die Darstellung überflutet, da sich nun nicht nur die Kanten eines Graphen, sondern Knoten und Kanten mehrerer Graphen überlagern. Strukturelle Besonderheiten sind dann kaum noch erkennbar und der Effekt wird umso stärker, desto dichter der Graph ist. Weiterhin werden verschiedene Farben verwendet, um die Zeitschritte unterscheiden zu können, weshalb keine Farben innerhalb des Graphen verwendet werden können, um Gruppenzugehörigkeiten zu markieren. Die Variante (c) ist vermutlich am ehesten verwendbar, da weder der Blickwinkel, noch die Überlagerung vorhanden sind. Der Vergleich von Positionen zwischen verschiedenen Zeitebenen ist jedoch deutlich schwieriger als zum Beispiel in (b), was umso schwieriger wird, desto weiter die Zeitschritte auseinander liegen.

3.4.2. Animation

Anstatt wie in [ITK10] jeden Zeitschritt als einzelne Ebene darzustellen, wird in [KG06] die tatsächliche Zeit als dritte Dimension verwendet, indem die einzelnen Zeitschritte als Animation nacheinander abgespielt werden. Dieser Ansatz vereinigt die in Abbildung 3.5 gezeigten Schritte (b) und (c), da die Veränderung von Positionen direkt ersichtlich wird und trotzdem jede Ebene einzeln sichtbar ist, wodurch die übermäßige Überlagerung der Darstellung wegfällt.

Da in [RFF+08] gezeigt wird, dass Animationen zwar schön anzusehen sind, aber häufig zu Fehlern bei der Lesbarkeit durch Nutzer führen, kann die Animation wie in [BC03] statisch angezeigt werden, indem die Graphen der einzelnen Zeitschritte dreidimensional übereinandergelegt werden, aber jede Ebene umso transparenter dargestellt wird, desto weiter sie vom aktuellen Zeitschritt entfernt ist. In Abbildung 3.6 ist ersichtlich, dass sowohl das Netzwerk des aktuellen Zeitschritts aufgrund der gesättigten Farben erkenntlich ist, als auch Veränderungen in der Existenz von Kanten gut nachvollzogen werden können. Allerdings führt die schnelle Zunahme der Transparenz dazu, dass nur wenige Zeitschritte in die Vergangenheit geschaut werden kann und der spezielle Vergleich von zwei bestimmten Zeitpunkten kaum möglich ist. Weiterhin wird die Darstellung umso schlechter lesbar, desto dichter der Graph ist.

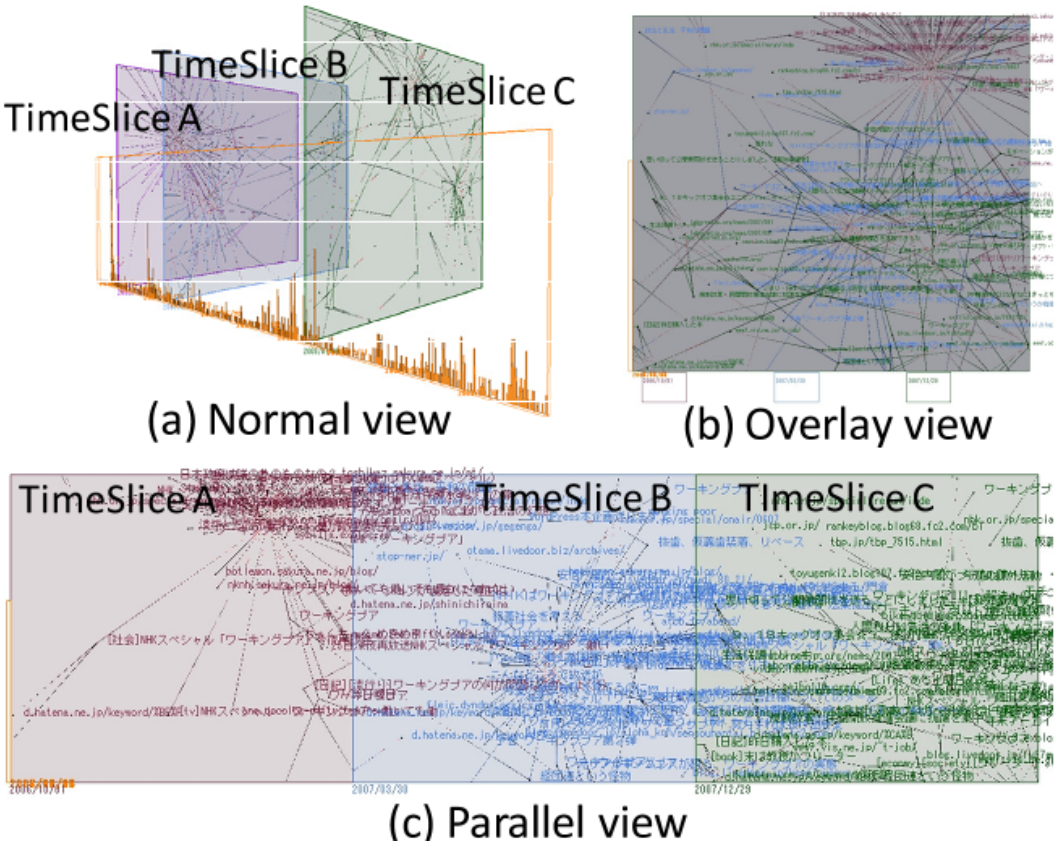


Abbildung 3.5.: Varianten der Visualisierung von TimeSlices [ITK10]. Die Daten werden nach ihrer zeitlichen Information gruppiert, jeder Gruppe eine Farbe zugeordnet und für jeden Zeitschritt ein Graph erstellt. In (a) werden die 2D-Graphen an einer Zeitleiste ausgerichtet, sodass eine 3D-Visualisierung entsteht. In (b) wird der Blick auf die 3D-Darstellung so festgelegt, dass nur noch die x-y-Ebene (Knotenpositionen) sichtbar ist, wodurch alle Zeitschritte transparent übereinander liegen. In (c) werden alle 2D-Graphen nebeneinander gelegt, wodurch die perspektivischen Fehler der 3D-Darstellung verschwinden und dennoch alle Graphen eingesehen werden können.

3.4.3. Zeitbereiche

Um die häufigen Probleme von dreidimensionalen Visualisierungen zu vermeiden, kann die Position der Knoten auf die Bewegung in einer Achse eingeschränkt werden,

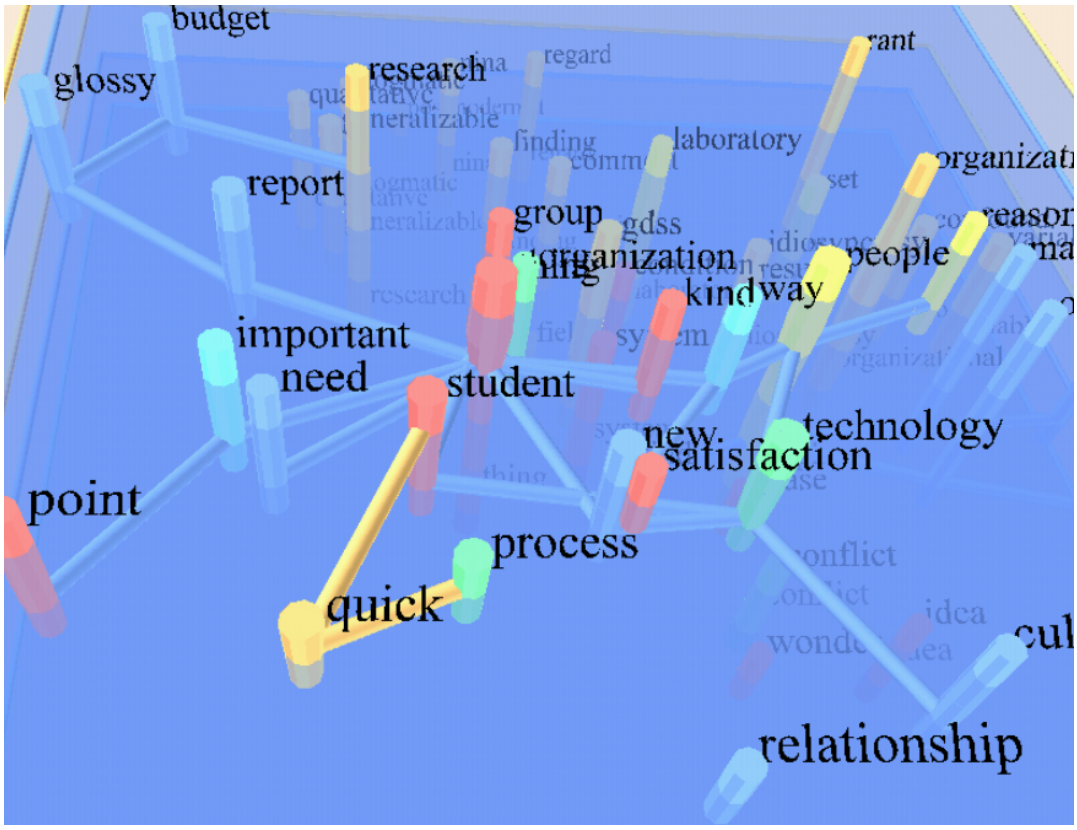


Abbildung 3.6.: Dynamic Centering Resonance Analysis [BC03]. Die Daten werden nach ihrer zeitlichen Information gruppiert und für jeden Zeitschritt wird ein Graph erstellt. Die zeitlichen Ebenen werden in der dritten Dimension übereinander gelegt und ihre Transparenz nimmt mit der zeitlichen Entfernung zum aktuellen Zeitpunkt zu. Der aktuelle Zeitschritt und die zeitliche Veränderung von Beziehungen ist für wenige Zeitschritte und dünn besiedelte Graphen möglich.

um die zweite Achse als Zeitachse zu verwenden. Alternativ kann die Position derart eingeschränkt werden, dass sich die Knoten nicht mehr in der kompletten Darstellung, sondern nur in einem für sie festgelegten Bereich bewegen dürfen. Dadurch kann der Platz der Visualisierung derart aufgegliedert werden, dass für jeden Zeitschritt ein Abschnitt zur Verfügung steht, wobei die Abschnitte unterschiedlich groß sein können, aber die Zeitpunkte noch immer sortiert sind. Solche Anwendungen finden sich unter anderem in den Publikationen [vEW14a] und [MGF12]. Die Prinzipielle

Struktur einer solchen Darstellung wird in Abbildung 3.7 dargestellt.

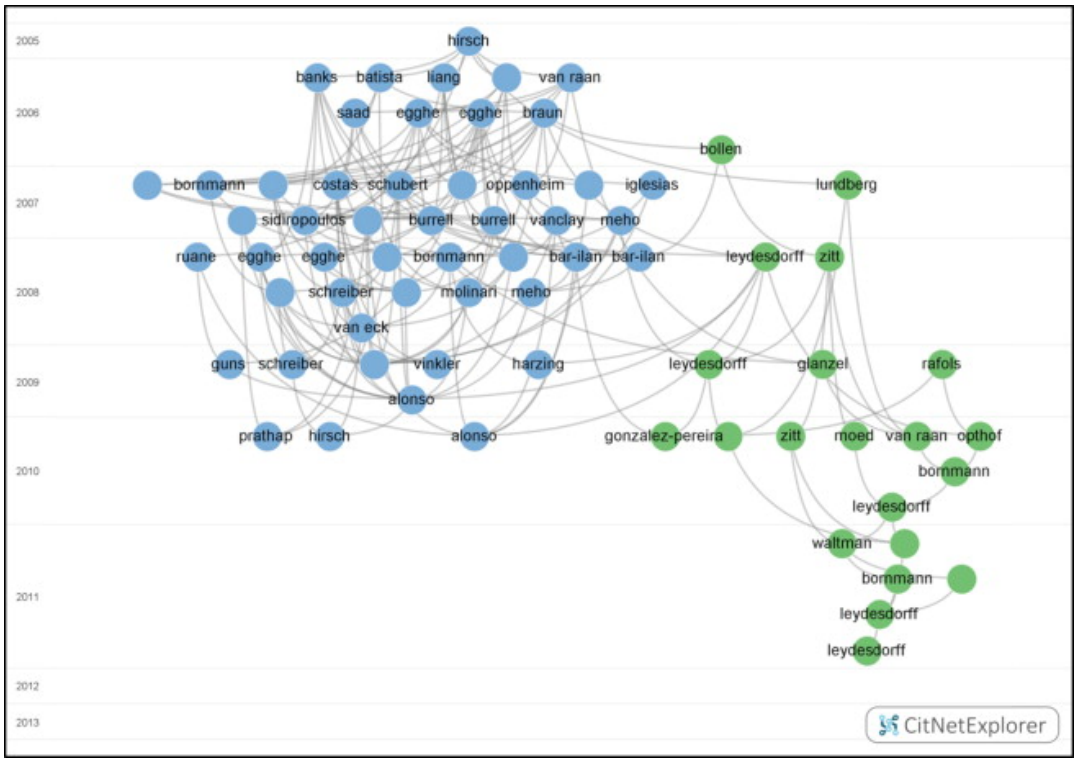


Abbildung 3.7.: CitNetExplorer [vEW14a]. Die Knoten werden nach ihrer zeitlichen Information gruppiert und entsprechend der für die Zeitschritte reservierten Bereiche angeordnet.

Die von uns verwendeten Daten (Kapitel 2) weisen jeder Publikation ein Erscheinungsjahr zu, wodurch sie als Knoten an einer Zeitleiste ausgerichtet werden könnten. Die in Abschnitt 3.3.3 genutzten Beziehungen von Publikationen zu Autoren können an dieser Stelle jedoch nur schlecht verwendet werden, weil weder ein Autor, noch ein Institut einer Zeit zugeordnet werden kann. Stattdessen ist die Darstellung eher für Zitierungen geeignet, da diese eine direkte Beziehung zwischen Publikationen zu verschiedenen Zeitpunkten ausdrücken.

Verändert man jedoch die Netzstruktur und passt die Visualisierung an gewissen Punkten an, ist dennoch die prinzipielle Idee anwendbar. Beispielsweise kann jede Publikation mit den Publikationen verbunden werden, welche einen Autor gemeinsam haben, jedoch bedeutet die Angabe desselben Autors weder, dass diese thematisch verwand sind, noch lässt sich ein Rückschluss auf die Beziehungen zwischen

Instituten ziehen. Stattdessen könnten Publikationen verbunden werden, welche ein veröffentlichendes Institut gemeinsam haben, wodurch sich innerhalb eines Zeitschrittes der übliche Graph entwickelt, welcher Publikationen desselben Instituts gruppiert und Beziehungen zwischen Clustern aufzeigt. Da sich nach dieser Vorgehensweise ein Cluster pro Institut entwickeln würde und Publikationen von mehreren Instituten zu den Clustern gezogen würden, welche mehr Publikationen haben, kann der Graph vereinfacht werden, indem für jedes Institut ein Knoten erstellt wird und Publikationen mehrerer Institute, in Form einer Kante, die Beziehung zwischen diesen darstellen. Weiterhin sollten die Kanten zwischen Zeitschritten derart reduziert werden, dass in jedem Zeitschritt t nur Kanten zum vorherigen und nachfolgenden Zeitschritt $t-1$ und $t+1$ gezogen werden. Wären die Areale für die einzelnen Zeitschritte so groß, dass sich der jeweilige Graph komplett entfalten könnte, so wäre die Darstellung der in Abbildung 3.5(c) ähnlich, mit dem Unterschied, dass auch Kanten zwischen Knoten verschiedener Zeitschritte vorhanden wären. Dies führt jedoch dazu, dass einzelne Kanten im schlimmsten Fall durch zwei komplette Graphen verlaufen, wodurch ein großer Teil der Graphen überdeckt würde, wenn viele solcher Kanten existieren.

Wenn man alternativ die Publikationen ihren Instituten zuordnet und für jedes Institut zwischen zwei Zeitschritten einen Knoten erzeugt (in Abbildung 3.7 wäre dies auf den Trennlinien), so könnte der Bereich der Zeitschritte für die Darstellung der Beziehungen zwischen Instituten genutzt werden. In diesem Fall gäbe es stets eine Kante zwischen zwei Institutsknoten, wenn in dem entsprechenden Jahr eine Publikation in Kooperation zwischen diesen beiden Instituten entstand. Abbildung 3.8 zeigt eine solche Anwendung auf Basis unserer Publikationsdaten.

Da alle Knoten derselben Fakultät übereinander angeordnet sind, bilden sich gleichfarbige Flächen, wenn Institute häufig innerhalb ihrer Fakultät kooperieren, während interdisziplinäre Zusammenarbeiten durch Farbverläufe und häufig größere Winkel der Kanten auffallen.

Da jeder Knoten eine feste Position hat, können in dieser Art des Graphen keine Kräftewirkungen stattfinden, wodurch die automatische Bildung von Gruppen nicht mehr möglich ist. Stattdessen wird ein besserer Überblick über die zeitliche Entwicklung von Instituten und Beziehungen gegeben, weil im Gegensatz zur parallelen Darstellung von Graphen (siehe Abschnitt 3.5(c)) die Breite der Zeitbereiche frei wählbar ist und somit der Abstand der zu vergleichenden Informationen reduziert wird.

Quer verlaufende und sich überkreuzende Kanten lassen die Darstellung chaotisch wirken, können jedoch mithilfe von Kantenbündelung verbessert werden. Auch die Interaktion durch Selektion von einzelnen Instituten kann besseren Einblick in die Daten gewähren. Weiterhin kann die Verwendung von Transparenz bei Kanten in Abhängigkeit ihrer Gewichte helfen, die Wichtigkeit einzelner Kooperationen herauszuarbeiten. In Abbildung 3.9 sieht man zum Beispiel, dass, wie auch in Abbildung 3.3 erkannt werden konnte, die drei durch die Farben Orange, Dunkelblau und Grau dar-

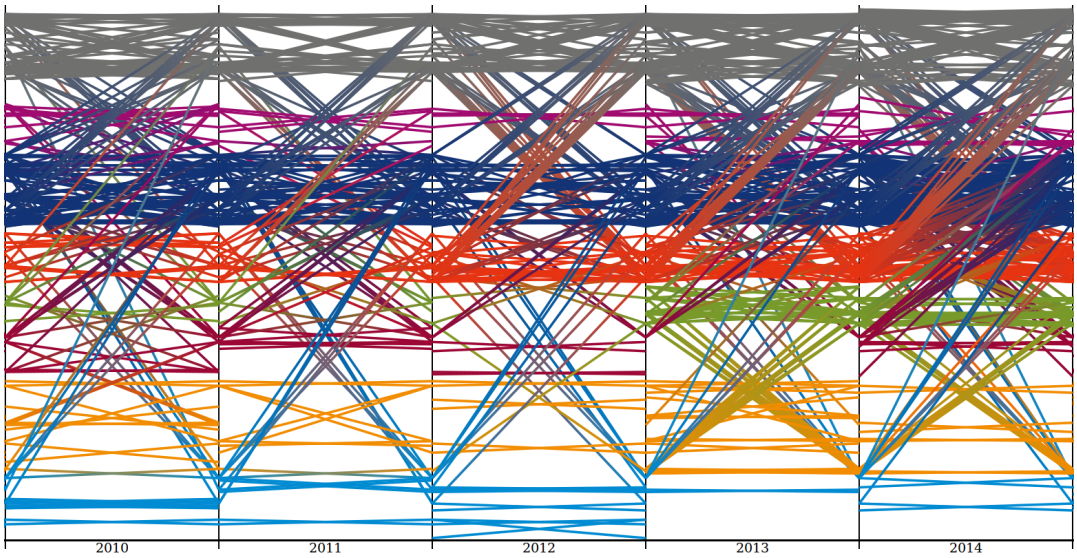


Abbildung 3.8.: Kooperationen zwischen Instituten der TU Chemnitz in den Jahren 2010 bis 2014. Jedem Zeitschritt wird ein Bereich in der Visualisierung zugeteilt, an dessen Anfang und Ende pro Institut (von unten nach oben sortiert nach ihrer Kostenstelle) ein Knoten in gleichmäßigem Abstand erstellt wird. Alle Knoten eines Instituts befinden sich auf derselben Höhe und sind nach dem Cooperate Design entsprechend ihrer Fakultätszugehörigkeit eingefärbt. Innerhalb eines Zeitabschnittes wird eine Kante zwischen zwei Institutsknoten erstellt, wenn in dem entsprechenden Jahr eine Kooperation zwischen den beiden stattfand. Die Kante verläuft dabei vom Institutsknoten zu Beginn des Zeitabschnittes zu dem Knoten des anderen Instituts am Ende des Zeitabschnittes. Für jede Publikation werden somit zwei Kanten erstellt, damit die Kooperation in beide Richtungen erkennbar ist. Jede Kante erhält einen linearen Farbverlauf entsprechend der Knoten mit denen sie verbunden ist.

gestellten Fakultäten den Größten Einfluss in der Universität haben. Dabei scheinen die Kooperationen zwischen Orange und Grau, sowie Blau und Grau jeweils stärker zu sein als die zwischen Blau und Orange. Weiterhin zeigt sich im Jahr 2013 eine deutliche Aktivität der gelben und grünen Fakultät, sowie eine starke Kooperation zwischen diesen beiden. Die Deckkraft d_i einer jeden Kante i wird dabei anhand der

Formel

$$d_i = \sqrt{\frac{w_i}{w_{max}}} \tag{3.5}$$

berechnet, wobei w_{max} das größte Kantengewicht aller gezeichneten Kanten ist.

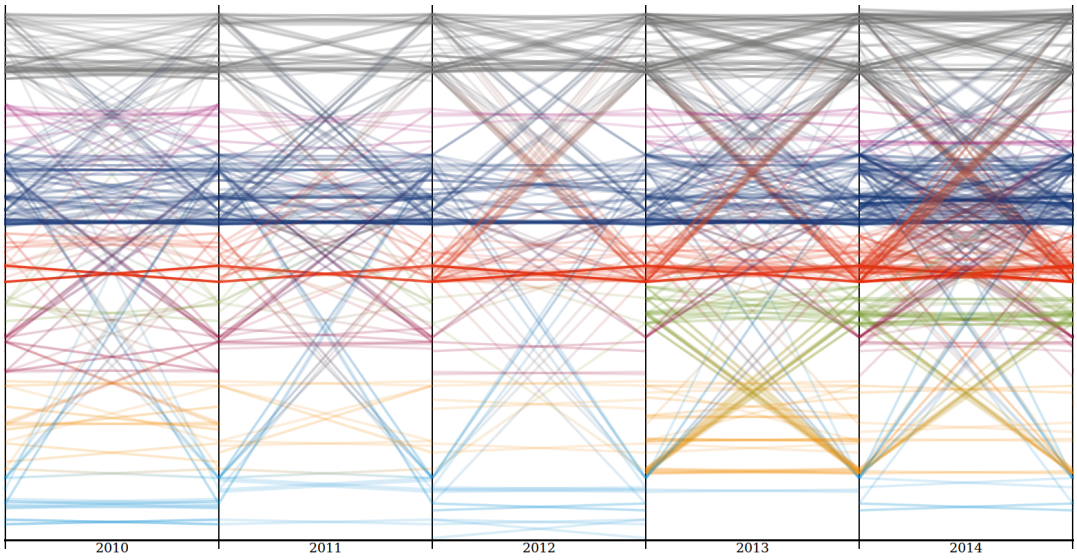


Abbildung 3.9.: Gewichtete Kooperationen zwischen Instituten der TU Chemnitz in den Jahren 2010 bis 2014. Die Kanten wurden genauso erstellt wie in Abbildung 3.8, mit dem Unterschied, dass Kanten mit niedrigeren Gewichten mit höherer Transparenz versehen wurden. Dazu wurde aus allen Kanten das höchste Kantengewicht ermittelt und die Deckkraft einer jeden Kante mit dem Faktor d_i multipliziert.

Möchte man die Kooperationen zwischen Instituten anteilhaft mit den publizierten Arbeiten innerhalb des Instituts vergleichen, müssten in jedem Zeitschritt Kanten von einem Institut zu sich selbst existieren, wodurch die komplette Darstellung farbig wäre, vorausgesetzt, dass jedes Institut jedes Jahr mindestens eine Publikation hatte. Die Menge an dann vorhanden Kanten würde auch durch Edge Bundling nicht dermaßen reduziert werden können, dass die Darstellung gut lesbar wäre. Auch das Hinzufügen von Kooperationen mit externen Mitarbeitern würde die Darstellung überfüllen, wenn ein Knoten *extern* hinzugefügt würde, weil dieser Knoten Kanten zu nahezu allen Instituten hätte.

3.5. Streamgraph

3.5.1. Einleitung

Es wurde gezeigt, dass Graphen gut zur Darstellung von Beziehungen zwischen verschiedenen Gruppen genutzt und einige nicht dargestellte Informationen unter Einschränkungen hinzugefügt werden können. Spätestens bei der Visualisierung der zeitlichen Entwicklung zeigt sich jedoch, dass das Prinzip des Graphen bei der Erweiterung um eine dritte Dimension zu starken Nachteilen in der Lesbarkeit führt. Verzichtet man auf die Position der Knoten und zeigt nur deren Beziehungen an, so wirkt die Darstellung überfüllt und chaotisch, insbesondere, wenn die Kooperationen mit internen, oder externen Arbeiten verglichen werden sollen.

Anstatt für die Menge an darzustellenden Informationen immer komplexere Visualisierungen zu nutzen, werden an dieser Stelle die Informationen derart reduziert, dass nicht mehr die Beziehung zwischen zwei speziellen Instituten, sondern allgemein die Anzahl der Kooperationen eines Instituts mit allen anderen Instituten betrachtet wird. Die Arbeiten werden dabei, wie in Kapitel 2.5 beschrieben, nach internen, intrafakultären, interfakultären und externen Publikationen unterteilt. Somit erhält man anstatt $O(n^2 \cdot t)$ möglichen Kanten nur noch $4 \cdot n \cdot t$ viele Zahlen, wobei n die Anzahl an Instituten und t die Anzahl betrachteter Zeitschritte ist.

Nach der Reduzierung der Datenmenge lässt sich unser Netzwerk mithilfe eines Streamgraphen darstellen, welcher pro Institut und Zeitschritt die Menge an Veröffentlichungen anzeigt und dabei sowohl einen allgemeinen, als auch einen detaillierten Überblick über die Daten liefert. Mit entsprechenden Erweiterungen und Interaktionsmöglichkeiten können alle in Kapitel 1.2 geforderten Zielstellungen erfüllt werden.

3.5.2. Grundidee

Für die Erläuterung der Erstellung eines Streamgraphen werden die in [BW08] beschriebenen Algorithmen verwendet. Für jedes Institut existiert pro Zeitschritt ein reeller, nicht-negativer Wert, welcher die Menge an Kooperationen in dem entsprechenden Jahr ausdrückt. Die Entwicklung des Wertes über die Zeit wird in [BW08] der Einfachheit halber als stetig differenzierbare Funktion f angegeben, welche im Intervall [0,1] definiert ist. Die Institute werden in eine feste Reihenfolge gebracht und ihre Werte entsprechend dieser Reihenfolge aufaddiert. Somit ergibt sich für jedes Institut eine neue Funktion g, welche ihre y-Position zum Zeitpunkt t entsprechend der Formel

$$g_i = g_0 + \sum_{j=1}^{i} f_j \tag{3.6}$$

angibt. Dabei ist i der Index des jeweiligen Instituts in der festgelegten Reihenfolge und g_0 der Wert der *Grundlinie* an der Stelle t, welche später erläutert wird.

Da in unserer Anwendung nur diskrete Werte zu festen Zeitpunkten vorhanden sind, werden die y-Positionen der Institute an den Zeitpunkten von der Funktion g abgelesen und zwischen den jeweiligen Punkte $(t, g_i(t))$ interpoliert. Die Grundlinie, Sortierung der Ebenen und die Interpolation der Punkte sind entscheidend für die resultierende Form des Streamgraphen, weshalb deren Variationen weiter erläutert werden.

Im einfachsten Fall ist $g_0 = 0$, wodurch der Graph auf der Zeitachse liegt (*Stacked Graph*) und die Interpolation erfolgt dermaßen, dass der Wert g_i im kompletten für den Zeitpunkt reservierten Intervall konstant ist und dem am entsprechenden Zeitpunkt definierten Wert entspricht (*Treppenfunktion*). Die Kombination wird in Abbildung 3.10 dargestellt.

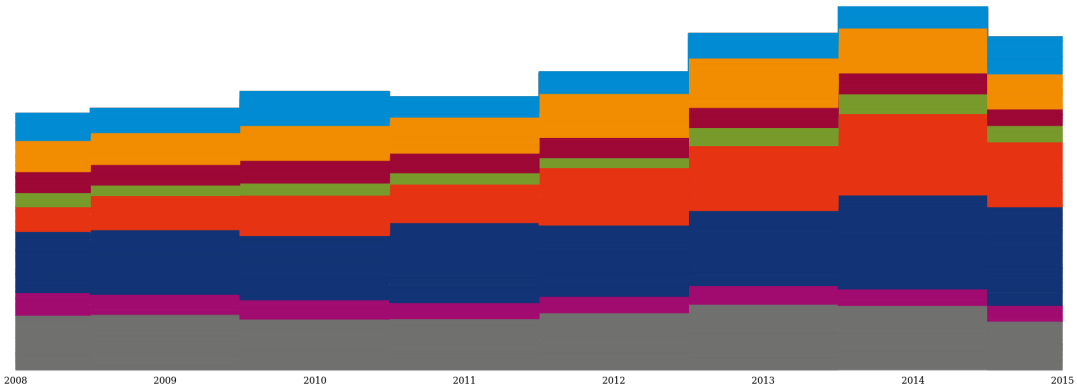


Abbildung 3.10.: Stacked Graph. Für jede Fakultät (entsprechend des Cooperate Design) wird die Anzahl an Publikationen pro Jahr angegeben und für jede Fakultät wird ihre y-Position anhand der Werte der unter ihr liegenden Fakultäten bestimmt. Dieser Wert wird im kompletten, für das Jahr reservierten, Zeitbereich konstant angezeigt, wodurch sich eine horizontale Linie pro Fakultät und Zeitbereich ergibt. Der y-Bereich zwischen dieser Linie und der darunterliegenden Linie kann für zusätzliche Informationen zum Institut verwendet und beispielsweise eingefärbt werden. Die Kombination aller y-Bereiche eines Instituts wird *Layer* oder Schicht genannt. Es lässt sich ablesen, dass von 2011 bis 2014 eine stetige Steigerung der Kooperationen an der TU Chemnitz stattfand. Die Steigerung lässt sich hauptsächlich auf die Entwicklung der orangen Fakultät zurückführen, da diese ebenso eine stetige Steigerung erfuhr, während alle anderen Fakultäten ungefähr gleichbleibende Werte hatten.

3.5.3. Interpolation

Die Interpolation der diskreten Punkte entscheidet sowohl über die Form des Graphen, und damit über die Ästhetik der Visualisierung, als auch über die Genauigkeit der ablesbaren Informationen. Neben der in Abbildung 3.10 gezeigten Treppenfunktion werden im Folgenden die stückweise lineare Interpolation und Cardinal Splines vorgestellt.

Die Treppenfunktion ist, besser als jede andere, dazu in der Lage, den Wert einer Schicht zu einem bestimmten Zeitpunkt darzustellen, da die Funktion nicht von angrenzenden Punkten beeinflusst wird. Des Weiteren zeigt sie dem Nutzer genau den für einen Zeitpunkt reservierten Bereich, weil sich über die Zeit verändernde Werte zu harten Kanten zwischen den Zeitbereichen entwickeln. Im Gegensatz zu interpolierten Werten kommt bei der Treppenfunktion eher der Bezug der Werte zu den Jahren zur Geltung, anstatt die Illusion zu schaffen, dass kontinuierliche Daten vorhanden seien.

Bei der stückweise linearen Interpolation wird zwischen allen Punkten P_n ein Streckenzug aufgebaut, welcher jeweils zwei zeitlich nebeneinanderliegende Punkte durch eine lineare Funktion verbindet. Die Strecke zwischen zwei Punkten P_i und P_j an den Zeitpunkten t und $t+1$ kann mithilfe der Formel

$$y = \frac{\Delta y}{\Delta x}x + \left(y_i - t\frac{\Delta y}{\Delta x}\right) = \frac{y_j - y_i}{x_j - x_i}x + \left(y_i - t\frac{y_j - y_i}{x_j - x_i}\right) \tag{3.7}$$

im Intervall $[x_i, x_j]$ dargestellt werden. Diese Art der Interpolation hat den Vorteil, dass im Gegensatz zur Treppenfunktion stets die Veränderung des Wertes an der Monotonie der Funktion ablesbar ist, wodurch die Entwicklung der Werte nicht durch Vergleiche hergestellt, sondern direkt visualisiert wird. Dies trifft insbesondere dann zu, wenn die Definition der Grundlinie dafür sorgte, dass die Schichten nahezu unabhängig voneinander sind. Da für die Erstellung des Streckenzugs direkt die vorhandenen Punkt verwendet werden, ist die Höhe der Schichten an den festgelegten Zeitpunkten immer korrekt. Die lineare Interpolation wird in Abbildung 3.11 gezeigt.

Anstatt lineare Funktionen für die stückweise Interpolation zu nutzen, können auch Funktionen höheren Grades verwendet werden. Um eine sowohl ästhetische, als auch genaue Darstellung zu erhalten, werden *Cardinal Splines* verwendet, welche stückweise Polynomfunktionen des 3. Grades entsprechend der Gleichung

$$p(t) = (2t^3 - 3t^2 + 1)p_0 + (t^3 - 2t^2 + t)m_0 + (-2t^3 + 3t^2)p_1 + (t^3 - t^2)m_1 \tag{3.8}$$

erstellen, wobei p_0 der Start- und p_1 der Endpunkt des Polynoms ist und $t \in [0, 1]$ gilt. Für die Berechnung von $n-1$ Polynomen zwischen n Punkten wird die jeweilige Tangente m_k am Punkt p_k mithilfe des vorherigen Punktes p_{k-1} und des nachfolgenden

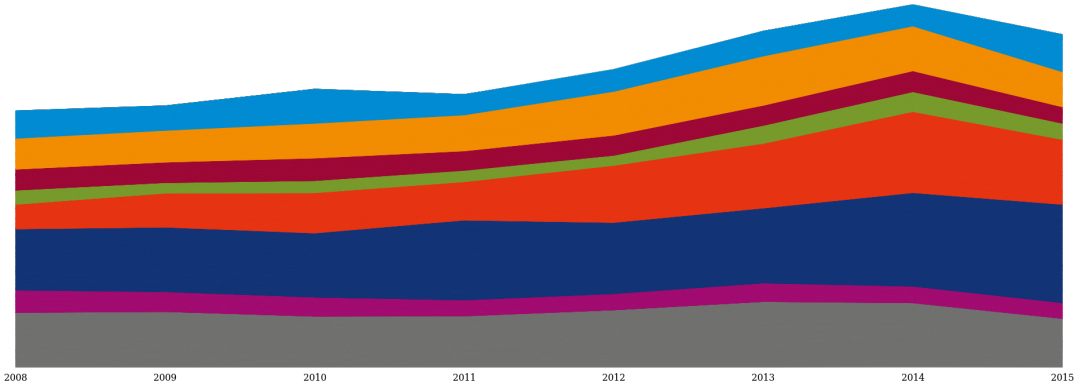


Abbildung 3.11.: Linear interpolierter Stacked Graph. Die für den Graphen nötigen Punkte werden genauso angeordnet wie in Abbildung 3.10, mit dem Unterschied, dass die Interpolation der Punkte anhand der Formel 3.7 erfolgt.

Punktes p_{k+1} entsprechend der Formel

$$m_k = (1 - c)\frac{p_{k+1} - p_{k-1}}{t_{k+1} - t_{k-1}} \tag{3.9}$$

berechnet. Die Konstante c gibt die *Spannung* an und liegt im Intervall $[0, 1]$. Wird $c = 1$ gesetzt, sind alle Anstiege an den Punkten 0, sodass die berechneten Polynome waagerecht durch die Punkte verlaufen. Setzt man $c = 0$, so entspricht der Anstieg dem einer Geraden, welche durch die Punkte p_{k-1} und p_{k+1} verläuft.

Cardinal Splines verlaufen, genau wie die Treppenfunktion und lineare Funktionen, exakt durch die Datenpunkte und geben somit exakte Auskunft über die Werte an den diskreten Zeitpunkten, weshalb sie besser zur Datenvisualisierung geeignet sind, als zum Beispiel B-Splines. Ähnlich wie bei den vorgestellten Streckenzügen kann anhand des Anstieges der Funktionen die Entwicklung des Wertes über die Zeit abgelesen werden, jedoch nimmt für den Nutzer die Illusion zu, kontinuierliche Daten zu betrachten. Durch die glatten Übergänge der Polynome wirkt die Darstellung jedoch deutlich ästhetischer.

3.5.4. Grundlinie

Definiert man die Grundlinie (*baseline*) $g_0 = 0$, erhält man einen Stacked Graph wie in Abbildungen 3.11, welcher insbesondere dafür geeignet ist, die globale Entwicklung des Netzwerkes aufzuzeigen, da für jedes Jahr die Summe aller Werte abgelesen

werden kann. Definiert man die Grundlinie stattdessen durch eine Funktion, lassen sich verschiedene Aspekte der Visualisierung optimieren. Im ThemeRiver [HHN00] wurde die Grundlinie als

$$g_0 = -\frac{1}{2}\sum_{i=1}^{n} f_i \tag{3.10}$$

definiert, wodurch $g_0 + g_n = 0$ gilt und die Silhouette des Graphen eine Symmetrie an der x-Achse aufzeigt. Im Vergleich zum Stacked Graph werden äußere Ausläufer weniger spitz dargestellt, weil die Werte nun auf beide Seiten der x-Achse aufgeteilt werden.

Der Begriff *wiggle* beschreibt, wie stark sich eine Schicht über die Zeit verändert. Desto größer die Veränderung einer Schicht ist, desto stärker wirkt sich dies auf die Position der darüber liegenden Schichten und ihrer zeitlichen Veränderung aus. Ein hoher wiggle führt dazu, dass auch Schichten ohne Änderung der Werte eine deutliche Änderung ihrer Position erfahren, wodurch ihre tatsächliche Entwicklung umso schlechter ablesbar ist. Die Minimierung des wiggle ist somit ein Kriterium zur Optimierung der Visualisierung und kann in die Definition der Grundlinie eingebracht werden. Die Formel

$$g_0 = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{i} f'_j \tag{3.11}$$

beschreibt, wie der wiggle des kompletten Graphen berechnet wird, indem für jede Schicht der Anstieg anhand des sich aufsummierenden Anstiegs der unter ihr liegenden Schichten berechnet und durch die Anzahl der vorhandenen Schichten geteilt wird, um den Durchschnitt zu ermitteln. Im diskreten Fall kann die Ableitung an einem Punkt P_t beispielsweise anhand des Anstieges der linearen Gleichung zwischen P_t und P_{t-1} berechnet werden.

Da dickere Schichten auch eine größere Veränderung erfahren können und häufig von Bedeutung für den Nutzer sind, sollte die Berechnung des wiggle die Dicke der Schichten mit einbeziehen. Die folgende Formel beschreibt eine gewichtete Variante der Grundlinienberechnung zur Minimierung des wiggle, wobei nun der wiggle in der Mitte einer jeden Schicht betrachtet wird.

$$g_0 = -\frac{1}{\sum_{i=1}^{n} f_i}\sum_{i=1}^{n} f_i \left(\frac{1}{2} f'_i + \sum_{j=1}^{i-1} f'_j\right) \tag{3.12}$$

Diese Definition der Grundlinie erleichtert die durchschnittliche Lesbarkeit einzelner Schichten zu allen Zeitpunkten, aber verschlechtert die Lesbarkeit der Silhouette des Graphen ein wenig. [BW08]

In der Implementierung von [BOH11] wird eine weitere Variante *expand* vorgestellt, welche die Grundlinie auf $g_0 = 0$ setzt, aber den Graphen in y-Richtung zu jedem

Zeitpunkt auf das Intervall $[0, 1]$ skaliert. Die Berechnung der einzelnen Schichten ändert sich dabei zu

$$g_i = \frac{\sum_{j=1}^{i} f_j}{\sum_{j=1}^{n} f_j} \tag{3.13}$$

. Abbildung 3.12 zeigt die Darstellung dieses Algorithmus.

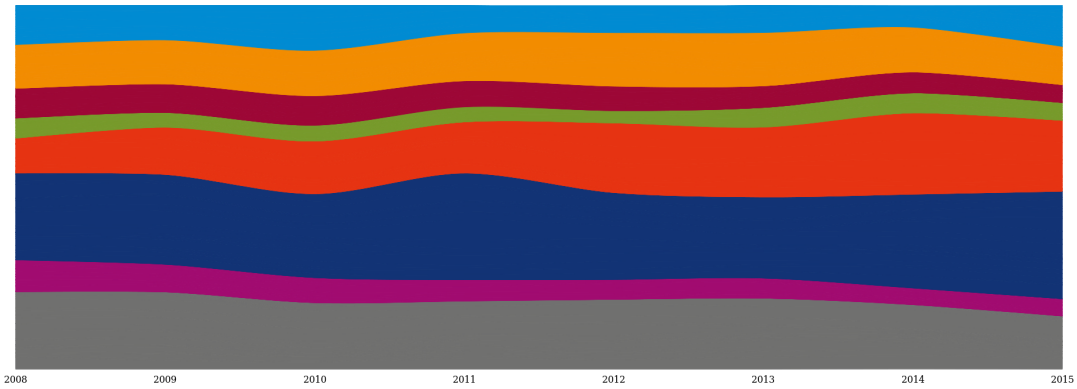


Abbildung 3.12.: Expand Streamgraph. Indem alle Werte zu jedem Zeitpunkt auf das Intervall [0,1] skaliert werden, wird der komplette für die Darstellung bereitgestellte Raum ausgenutzt. In dieser Ansicht lassen sich insbesondere prozentuale Anteile einzelner Schichten von den Gesamtwerten gut ablesen und mit denen anderer Schichten vergleichen.

3.5.5. Sortierung

Die Sortierung der Schichten ist ein weiteres Kriterium, welches die Darstellung beeinflusst und die Geschwindigkeit im Finden von gesuchten Informationen beschleunigen kann. In [BW08] wird gezeigt, dass die Sortierung nach der Startzeit einer Schicht (Zeitpunkt, an dem die Schicht zum ersten Mal einen Wert ungleich null aufweist) zu stark asymmetrischen Graphen führt, wenn alle Schichten derart geformt sind, dass sie mit einer starken Steigerung beginnen und ihre Werte anschließend kontinuierlich sinken. Für eine solche Datenlage wurde der *inside-out*-Algorithmus vorgestellt, welcher den Graphen in eine untere und obere Hälfte teilt, die Schichten aufsteigend nach ihrer Startzeit sortiert und anschließend nacheinander die Schichten jeweils zu der Hälfte hinzufügt, welche aktuell die geringere Dicke hat. Der Algorithmus führt zu einem eher symmetrischen Bild, da das Hinzufügen einer dicken Schicht zu einer Hälfte auf der anderen Seite ausgeglichen wird, indem diese so lange Schichten

hinzugefügt werden, bis dieselbe Dicke erreicht oder überschritten wurde. Da unserer Datenbasis erst ab einem Zeitpunkt betrachtet wird, ab dem fast alle Institute begonnen haben, veröffentlichte Arbeiten in die Datenbank einzutragen, würde eine solche Sortierung nur dafür sorgen, dass neu gegründete Institute an den Rand der Darstellung verschoben werden. Des Weiteren werden Institute eher mit der Zeit größer, anstatt zeitig ihr Maximum zu erreichen und dann abzuschwächen, wodurch Institute, welche bereits lange Zeit bestehen in der Mitte der Darstellung positioniert und aufgrund ihrer Dicke eine starke Verschiebung der anderen Schichten verursachen würden.

In [DBH16] wird ein allgemeinerer Ansatz gewählt, welcher von der Datenbasis unabhängig ist und versucht, zusätzlich zur Grundlinie, auch mit der Sortierung den wiggle des Graphen zu minimieren. Dazu startet der Algorithmus mit einer horizontalen Linie und definiert die obere und unter Hälfte. Anschließend wird jede noch nicht hinzugefügte Schicht einzeln darauf getestet, wie sich der wiggle des Graphen entwickelt, wenn sie zu der oberen oder unteren Hälfte hinzugefügt wird. Die Schicht, welche beim Hinzufügen zu einer der beiden Hälften den kleinsten wiggle verursacht, wird schließlich auf dieser Seite hinzugefügt. Der Vorgang wird für jede nicht hinzugefügte Schicht wiederholt. Der Algorithmus wird weiterhin verbessert, indem im Anschluss stets zwei benachbarte Schichten iterativ darauf getestet werden, ob ihre aktuelle Reihenfolge einen größeren wiggle erzeugt, als wenn die beiden vertauscht würden.

In Kapitel 4.7 werden weitere Sortierverfahren gezeigt, welche auf unsere Daten und die Zwecke der Anwendung angepasst sind.

4. Implementierung

4.1. Einleitung

In Kapitel 3 wurde gezeigt, dass Graphen ungeeignet sind, um die Entwicklung von Beziehungen zwischen Instituten über die Zeit darzustellen. Aus diesem Grund wurden die Daten von speziellen auf allgemeine Kooperationen vereinfacht und können mithilfe eines Streamgraphen dargestellt werden. Durch die Implementierung einiger Erweiterungen und Interaktionen wird der Streamgraph dazu befähigt, Institute und Fakultäten in Hinblick auf ihre Kooperationen hin zu untersuchen und dabei auch das Verhältnis verschiedener Kooperationstypen darzustellen.

Damit die Anwendung einer breiten Masse von Nutzern zur Verfügung gestellt werden kann, basiert die Implementierung des Streamgraphen auf den Webtechnologien HTML, CSS und JavaScript, sowie den JavaScript-Bibliotheken D3 [BOH11] in der Version 3.5.17 und jQuery in der Version 3.1. Die Berechnung der Kooperationsdaten aus der Datenbank nach JSON erfolgt mittels PHP.

4.2. Überblick

Der für diese Anwendung erstellte Streamgraph basiert auf den in Kapitel 2.6 beschriebenen, extrahierten Daten, welche derart gefiltert wurden, dass nur Institute betrachtet werden, deren Kostenstellen mit der Ziffer 2 beginnen, womit sie zum akademischen Bereich gehören. Weiterhin wurden alle Einträge vor dem Jahr 2008 entfernt, sowie das aktuelle Jahr 2016, da für dieses noch kein vollständiger Datensatz vorliegt und einige Institute alle Publikationen des Jahres zu einem Termin in die Datenbank eintragen.

Die Standardwerte zur Berechnung des Graphen sind so gewählt, dass alle vier Kooperationskategorien ihren Wert zur Dicke der Schichten beisteuern, alle Schichten nach ihrer Kostenstelle sortiert sind, wodurch Institute derselben Fakultät nebeneinander liegen, die Grundlinie anhand des *wiggle*-Algorithmus aus Kapitel 3.5.4 berechnet wird und die Interpolation mithilfe von Cardinal Splines geschieht. Die Farben der Institute entsprechen denen des Cooperate Design [coo14], womit alle Institute derselben Fakultät einheitlich eingefärbt sind.

Der Streamgraph bietet viele Ansatzpunkte zur Interaktion, die das Aussehen des Graphen verändern und unterschiedliche Informationen preisgeben. So lassen sich nicht nur die für die Berechnung des Streamgraphen verwendeten Algorithmen

zur Grundlinie und Interpolation ändern, sondern auch Selektion, Filterung, Zoom-, Scroll- und Suchfunktionen umsetzen, die die Exploration des Graphen erleichtern. Weiterhin wird eine Option zur Auswahl der angezeigten Kooperationsarten bereitgestellt und ein Tooltip, welcher weitere Informationen zu markierten Schichten liefert. Schließlich wird ein Gradient zur Verfügung gestellt, welcher den Vergleich verschiedener Kooperationsarten zueinander ermöglicht. Im Folgenden werden die implementierten Interaktionsmöglichkeiten genauer beschrieben. Die Menüs zur Steuerung der Interaktionen werden in Abbildung 4.1 aufgezeigt.

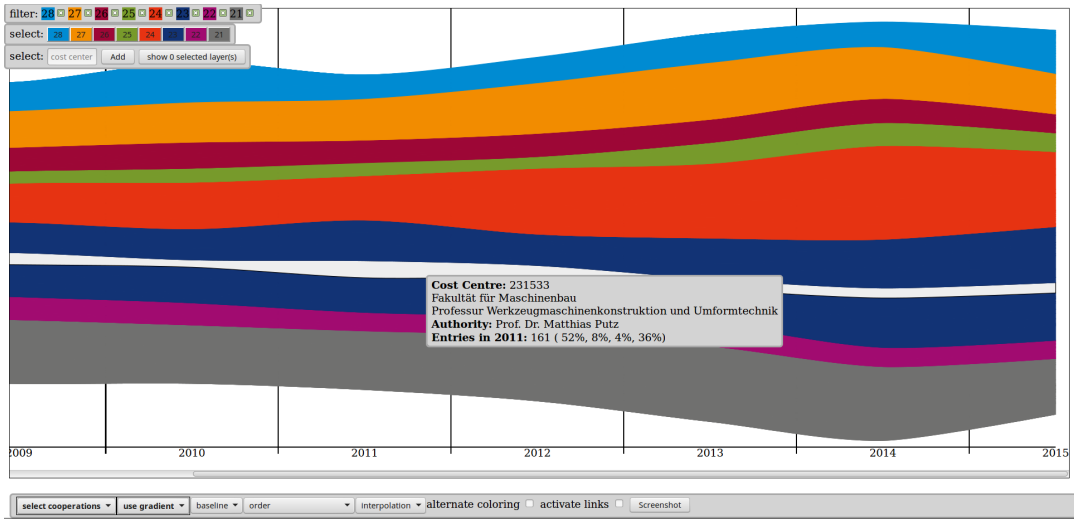


Abbildung 4.1.: Implementierung. Die Anwendung verwendet den größten Teil des gegebenen Platzes zur Darstellung des Streamgraphe. Eine Zeitachse markiert die jeweiligen Zeitbereiche und beschriftet diese. Entlang der Zeitachse kann zu vergangenen Jahren gescrollt werden. Ein Tooltip gibt zusätzliche Informationen zur markierten Schicht am ausgewählten Zeitpunkt. Oben links finden sich Optionen zur Filterung, Selektion und Suche. Unten werden die Optionen zur Datenänderung, Einfärbung, Sortierung und Berechnung des Graphen bereitgestellt.

4.3. Grundlinie und Interpolation

Die aus dem Kapitel 3.5.2 bekannten Algorithmen zur Berechnung des Streamgraphen werden dem Nutzer frei kombinierbar als Umschaltfunktion bereitgestellt, da diese unterschiedliche Vorteile besitzen.

Wird die Grundlinie $g0 = 0$ (*zero*) gesetzt, lassen sich leicht Gesamtwerte ablesen. Im Modus *expand* ist genau das Gegenteil der Fall, da die Summe aller Werte zu jedem Zeitpunkt einhundert Prozent entsprechen, kann keine Aussage über Gesamtwerte getroffen werden, aber der Vergleich von Schichten fällt deutlich leichter und auch Jahre mit niedrigen Gesamtwerten, und dadurch dünnen Schichten, können einfacher verglichen werden. Zusätzlich kann der prozentuale Anteil einer oder mehrerer Schichten von der Gesamtheit der angezeigten Schichten leicht bestimmt werden. Der Modus *wiggle* bietet eine Kombination aus beiden und versucht sowohl die Gesamtheit, als auch lokale Besonderheiten erforschbar zu machen.

Die Interpolation *cardinal* sorgt für eine ästhetische Form des Graphen und zeigt dennoch korrekte Werte an den definierten Zeitpunkten, allerdings erzeugt er auch die Illusion von kontinuierlich vorliegenden Daten. Die Treppenfunktion *step* zeigt deutlich, welcher Wert zu welchem Zeitbereich gehört und ist insbesondere dafür geeignet, möglichst korrekte Werte abzulesen. Die *polyline* hat im Vergleich zur Interpolation mit Cardinal Splines keinen wirklichen Vorteil, jedoch wird sie, wie in der später beschriebenen Evaluation herauskam, von einigen Nutzern bevorzugt genutzt.

4.4. Scroll- und Zoomfunktion

Scroll- und Zoomfunktionen sind zwei der grundlegendsten Interaktionen, die jeder Anwendung zur Exploration beigefügt sein sollten und beeinflussen sich häufig gegenseitig. Die Zoomstufe in x-Richtung bezeichnet das Verhältnis von der Anzahl der vorhanden x-Werte zu der Anzahl der angezeigten x-Werte. In der Standardeinstellung werden über die Breite der Anwendung verteilt sechs komplette Zeitbereiche von den insgesamt acht berechneten angezeigt, womit sich ein Zoomfaktor von $8/6 = 1.3\overline{3}$ ergibt. In Abhängigkeit des Zoomfaktors verändert sich auch die Breite des Scrollbalkens, welcher den inversen Zoomfaktor als Skalierungsfaktor von der für den Scrollbalken zur Verfügung stehenden Breite nutzt. In diesem Fall nimmt der Scrollbalken das $1.3\overline{3}^{-1} = 0.75$-fache des Platzes ein. Der minimale Zoomfaktor wird auf 1 gesetzt, da in diesem Fall die komplette Darstellung angezeigt wird und auch kein Scrollen mehr möglich ist. Auf die Zoom- und Scrollfunktionen in y-Richtung wurde in dieser Anwendung verzichtet, da die Schichten durch Selektion und Filterung vergrößert werden können.

(Anmerkung: Die in diesem Abschnitt berechneten Faktoren werden zusätzlich durch die in der Anwendung verwendeten Seitenränder der Darstellung beeinflusst. Des Weiteren werden für den ersten und letzten Zeitschritt nur jeweils halbe Zeitbereiche bereitgestellt, womit sich die Anzahl der darstellbaren Zeitbereiche auf $t - 1$ beschränkt.)

4.5. Datenauswahl

Die vier Kooperationskategorien, in welche die Werte eines jeden Instituts unterteilt wurden, können durch den Nutzer aktiviert und deaktiviert werden, sodass die Schichten nur so dick wie die Summe der aktivierten Kategorien sind. Der resultierende Wert y_r eines Instituts zum Zeitpunkt t wird anhand der Formel

$$y_r = a_0 y_{sameInst} + a_1 y_{sameFac} + a_2 y_{diffFac} + a_3 y_{extern} \tag{4.1}$$

berechnet, wobei $a_x = 1$ gilt, wenn die jeweilige Kategorie x aktiviert ist, ansonsten ist $a_x = 0$. Abbildung 4.2 zeigt den Streamgraphen, wenn ausschließlich Kooperationen mit anderen Fakultäten dargestellt werden. Dabei wird deutlich, dass in den Jahren 2012 bis 2014 ein starker Anstieg der interfakultären Publikationen stattfand und die Zahl dieser Kooperationen um mehr als das doppelte anstieg.

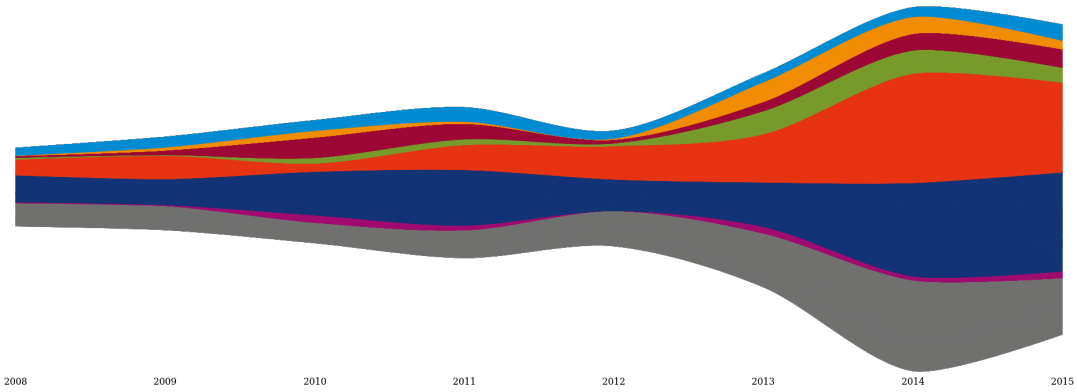


Abbildung 4.2.: Interfakultäre Kooperationen. Die Datenbasis kann durch den Nutzer derart verändert werden, dass nur noch Kooperationen zwischen verschiedenen Fakultäten gewertet werden.

4.6. Tooltip

Bewegt man den Mauszeiger über eine der angezeigt Schichten, so wird diese farblich markiert und ein Tooltip angezeigt, welcher zusätzliche Informationen enthält. Diese umfassen die Nummer der Kostenstelle, die Bezeichnung des Instituts, deren zugehörige Fakultät, den Namen der für das Institut zuständigen Person und den berechneten Wert y_r am Zeitpunkt des ausgewählten Zeitbereiches. Weiterhin wird in Prozentzahlen angegeben, wie sich der errechnete Wert auf die Kooperationskategorien aufteilt. Ein möglicher Tooltip wird in Abbildung 4.1 gezeigt. Der Tooltip ist

in unserer Anwendung für zwei Dinge wichtig: Zum einen hilft er dem Nutzer, aufzuzeigen, was dargestellt wird, indem er Informationen zu den Schichten bereithält, die der Anwender seinen bisherigen Kenntnissen zuordnen kann. So kann er zum Beispiel herausfinden, welche Farbe welche Fakultät repräsentiert, indem er über mehrere Schichten der gleichen Farbe fährt. Dies hilft weiterhin, die Gruppierung von Instituten und die Repräsentation einzelner Institute, durch die Schichten, zu verstehen. Zweitens zeigt der Tooltip konkrete Zahlen in Form des dargestellten Wertes und dessen prozentualer Aufteilung an, welche derart nirgends in der Darstellung zu finden ist. An dieser Stelle sei angemerkt, dass auf eine angezeigte y-Achse zur Angabe des Wertebereiches verzichtet wurde, um den berechneten Kooperationswert nicht erklären zu müssen. Wie bereits erwähnt spiegelt dieser nicht die Anzahl der tatsächlichen Publikationen in einem Jahr wider und soll auch nicht den Anschein erwecken, dass er dies tun würde.

4.7. Sortierung

In Kapitel 3.5.5 wurde gezeigt, dass die Sortierung nach dem Erstauftreten eines Instituts für unsere Daten nicht sinnvoll ist und die Sortierung zur Minimierung des wiggle, aufgrund der sich stark unterscheidenden Farben der Schichten, zu einer chaotisch wirkenden und verwirrenden Darstellung führen würde. Weiterhin ist dies eine Sortierung, deren Nutzen für den Anwender nicht direkt ersichtlich ist und deshalb eher im Hintergrund agieren sollte, anstatt als Option wählbar zu sein.

Standardmäßig erfolgt die Sortierung nach der Kostenstelle, wodurch alle Institute einer Fakultät aneinandergrenzen und aufgrund ihrer Einfarbigkeit auch Fakultäten vergleichbar machen, ohne diese als extra Strukturelement in die Visualisierung einzubauen. Mithilfe dieser Sortierung ist es für den Anwender besonders einfach, spezielle Institute zu finden, wenn er deren Kostenstelle kennt und selbst wenn nicht, müssen nur die Institute einer Fakultät mithilfe des Tooltips untersucht werden.

Da Anwender häufig an den Extremen einer Darstellung interessiert sind, kann die Sortierung nach dem Wert in einem bestimmten Jahr erfolgen, wodurch die Schicht mit dem höchsten Wert ganz oben gezeichnet wird und somit sofort zu finden ist. Eine solche Sortierung wird in Abbildung 4.3 vorgeführt, in welcher nebenbei auch das zuvor erwähnte Chaos der abwechselnden Farben zur Schau gestellt wird.

Anstatt die Werte aller Schichten miteinander zu vergleichen, können auch nur Schichten derselben Fakultät miteinander verglichen werden. Somit bleibt die Zugehörigkeit der Institute nicht nur anhand ihrer Farben, sondern auch anhand ihrer Positionen erhalten. Die Sortierung kann in Abbildung 4.4 betrachtet werden.

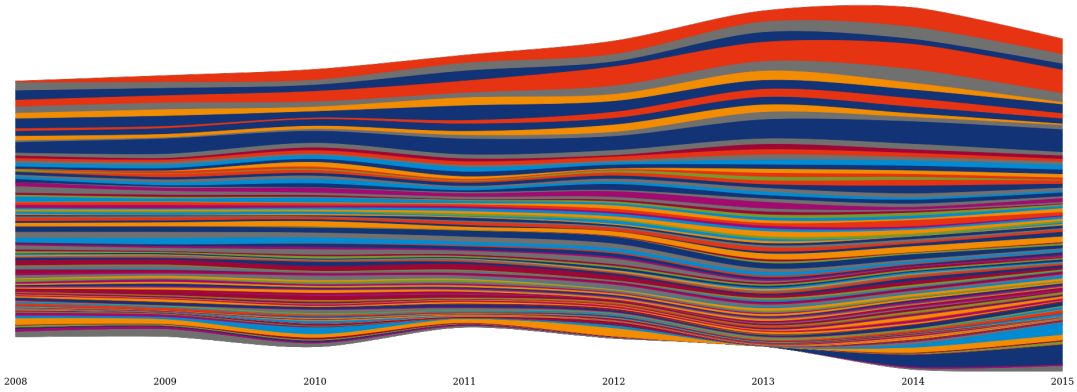


Abbildung 4.3.: SortByMaxAtTime. Die Sortierung der Schichten nach ihrem maximalen Wert im Jahr 2013 verschiebt die wichtigsten Schichten des gewählten Jahres an die Spitze der Darstellung. Obwohl es so aussieht, als ob die vierte Schicht von oben dicker sei, als die darüber liegenden, handelt es sich tatsächlich um zwei Schichten. Diesem Trugschluss lässt sich entgegenwirken, indem benachbarten Schichten derselben Fakultät verschiedene Farben zugeordnet werden. Lässt man die Farben jedoch so, kann man neben den lokalen Maxima auch einschätzen, welche Fakultät in den Top Ten des Jahres am häufigsten vorkommt.

4.8. Filterung, Selektion und Suche

Die Filterfunktion erlaubt es dem Nutzer in jeder Situation die Fakultäten auszublenden, welche für ihn nicht von Interesse sind.

Die Selektion ist das wichtigste Werkzeug dieser Anwendung, um den Fokus des Nutzers auf die Visualisierung zu übertragen. Sie ermöglicht nicht nur die genauere Untersuchung einzelner Schichten wie in Abbildung 4.5, sondern insbesondere die Auswahl einer beliebigen Anzahl von zuvor dargestellten Instituten und Fakultäten.

Abbildung 4.6 demonstriert, wie mehrere Institute und sogar ganze Fakultäten mit einem Mal ausgewählt werden können, um deren Verläufe zu vergleichen.

Die Selektion ersetzt in gewisser Weise die Zoomfunktion in y-Richung, weil sie ermöglicht, weniger Schichten auf dieselbe Höhe auszubreiten, wie zuvor die Gesamtheit aller Schichten. Zusätzlich vereinfacht sie den Vergleich ausgewählter Schichten, da ihr Abstand zueinander geringer wird, falls zuvor zwischen ihnen liegende Schichten nicht mit selektiert wurden.

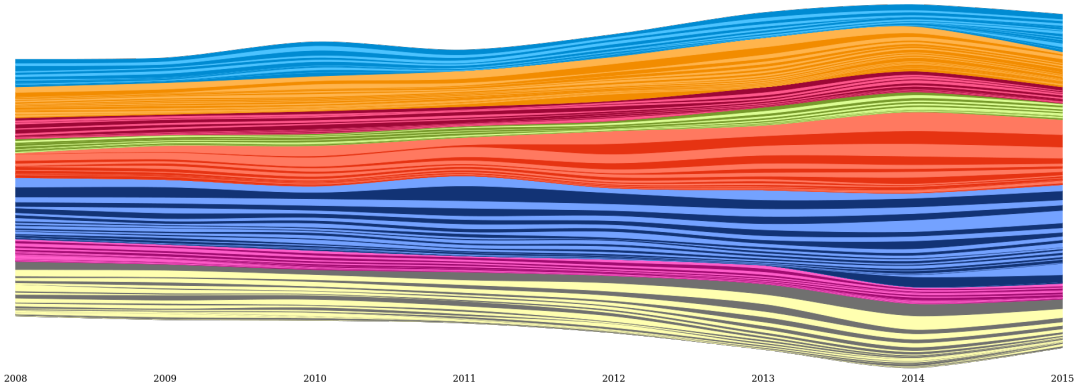


Abbildung 4.4.: SortByMaxInFacultyAtTime. Die Sortierung der Schichten innerhalb der Fakultät nach ihrem maximalen Wert im Jahr 2013 hält die positionelle Zuordnung der Schichten zu ihren Fakultäten aufrecht und erleichtert es, die wichtigsten Institute eines Jahres einer jeden Fakultät zu bestimmen. Das Alternieren der Farben nebeneinander liegender Schichten derselben Fakultät ermöglicht die Unterscheidung der einzelnen Institute voneinander.

4.9. Gradient

Während die *expand*-Funktion aus Kapitel 3.5.4 dazu geeignet ist, den prozentualen Wert einer Schicht vom Gesamtwert aller dargestellten Schichten abzulesen, dient die Gradientenfunktion dazu, den prozentualen Anteil der Bestandteile des Wertes einer Schicht vom Gesamtwert dieser Schicht abzuschätzen. Da der Wert einer jeden Schicht in unserer Anwendung aus den vier Kooperationskategorien berechnet wird, dient der hier berechnete Faktor y_{part} dazu, den Anteil ausgewählter Kategorien von den dargestellten Kategorien festzustellen. Zur Berechnung des Faktors dient die Formel

$$y_{part} = \frac{b_0 y_{sameInst} + b_1 y_{sameFac} + b_2 y_{diffFac} + b_3 y_{extern}}{a_0 y_{sameInst} + a_1 y_{sameFac} + a_2 y_{diffFac} + a_3 y_{extern}} \tag{4.2}$$

wobei $a_x = 1$ gilt, wenn die jeweilige Kategorie x aktiviert ist, ansonsten ist $a_x = 0$. Weiterhin gilt $b_x = 0$, wenn $a_x = 0$, oder die Kategorie x im Gradienten deaktiviert ist, ansonsten ist $b_x = 1$. Aus der Gleichung folgt, dass nur Anteile Einfluss auf den Faktor y_{part} haben können, die auch zur Berechnung des Wertes y_r aktiviert waren. Der Faktor y_{part} bestimmt die Deckkraft der für die Schicht verwendeten Farbe zum Zeitpunkt t. Je nach der Beschaffenheit der Werte der ausgewählten Schichten kann die Anwendung des Gradienten zu sehr blassen Farben führen. Um dem entgegenzu-

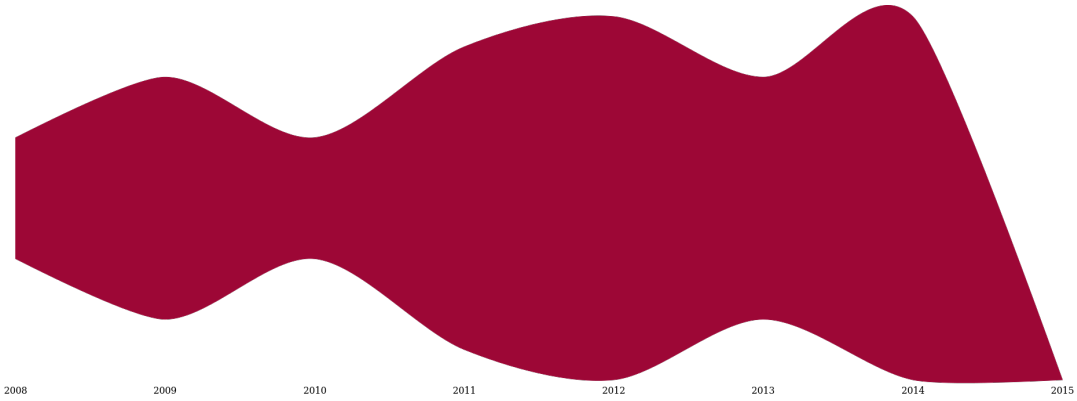
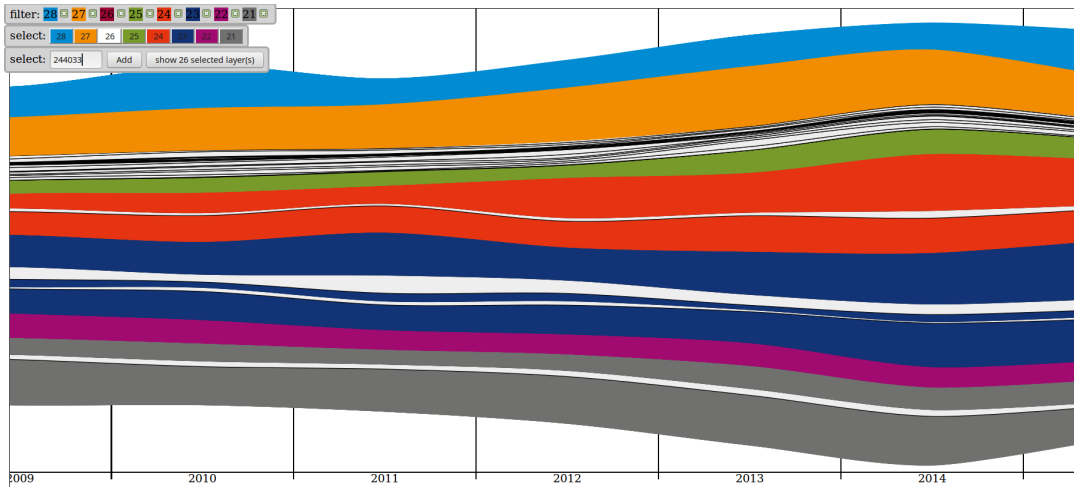


Abbildung 4.5.: Selektion einzelner Schichten. Die hier ausgewählte Schicht befand sich in der Sortierung nach dem Wert im Jahr 2013 in den unteren fünf Prozent der Darstellung, womit ihre Werte so gering sind, dass sie sich in der globalen Ansicht kaum untersuchen lassen. Indem nur die Werte dieser einen Schicht dargestellt und auf die Höhe der Anwendung skaliert werden, kann ihre zeitliche Entwicklung genauer betrachtet werden.

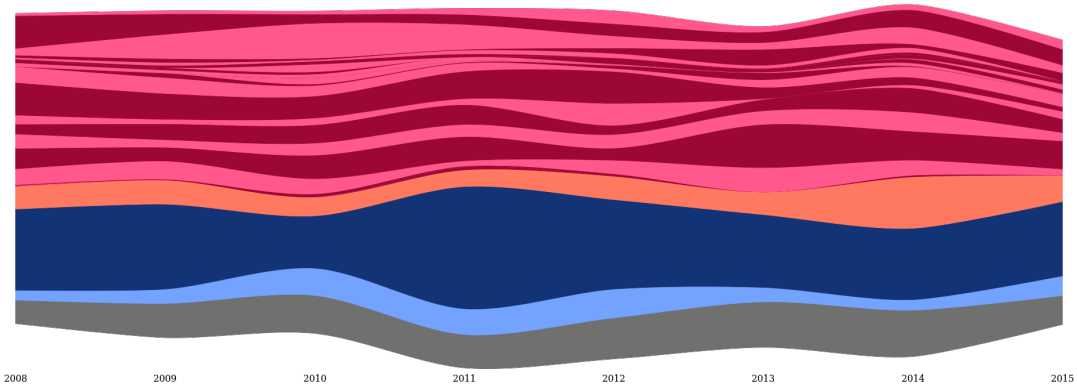
wirken können die Farben normiert werden, indem aus allen dargestellten Schichten der maximale, vorkommende Faktor $max(y_{part})$ herausgesucht wird und der Faktor y_{part} an jedem Zeitpunkt einer jeden Schicht durch $max(y_{part})$ dividiert wird. Da $y_{part} <= max(y_{part}) <= 1$ gilt, kann die Division nie zu einer Verminderung des Faktors führen. Wenn $max(y_{part}) = 0$ gilt, wird die Division nicht ausgeführt. Ein Beispiel zur Verwendung des Gradienten wird in Abbildung 4.7 demonstriert. Diese lenkt den Blick auf Institute, welche bei einem hohen Anteil ihrer Veröffentlichungen mit Instituten anderer Fakultäten zusammengearbeitet haben. Das Herauslesen einer solchen Information wäre ohne die Anwendung eines Gradienten nur mit sehr hohem Aufwand des Nutzers möglich.

4.10. Kooperationskanten

Nachdem beim Übergang vom Graphen aus Kapitel 3.3 zum Streamgraphen die Information der Kooperationen zwischen speziellen Instituten verallgemeinert wurde, können diese Beziehungen im Nachhinein wieder ergänzt werden, indem in jedem Zeitbereich jeweils eine Kante zwischen zwei Instituten verläuft, welche im entsprechenden Jahr miteinander kooperierten. Die resultierende Darstellung wird in Abbildung 4.8 gezeigt und ähnelt stark dem Vorgehen aus Abbildung 3.8 mit dem Unter-



(a) Einzelne Institute können per Klick, oder durch die Eingabe ihrer Kostenstelle in die Suchfunktion zur Auswahl hinzugefügt werden. Weiterhin wurden alle Institute der Fakultät ausgewählt, deren Kostenstelle mit der Nummer 26 beginnt. Alle aktuell ausgewählten Schichten werden markiert, indem ihre Farbe zu weiß wechselt und sie einen schwarzen Rand erhalten.



(b) Alle ausgewählten Schichten können anschließend in einem neuen Streamgraphen angezeigt werden. Um die Schichten gleicher Fakultäten unterscheiden zu können, wurden alternierende Farben verwendet.

Abbildung 4.6.: Selektion mehrerer Schichten.

schied, dass die Institute nicht alle im gleichen Abstand zueinander stehen, sondern der Abstand von der Dicke der jeweiligen Schichten abhängt. Der Startpunkt A und

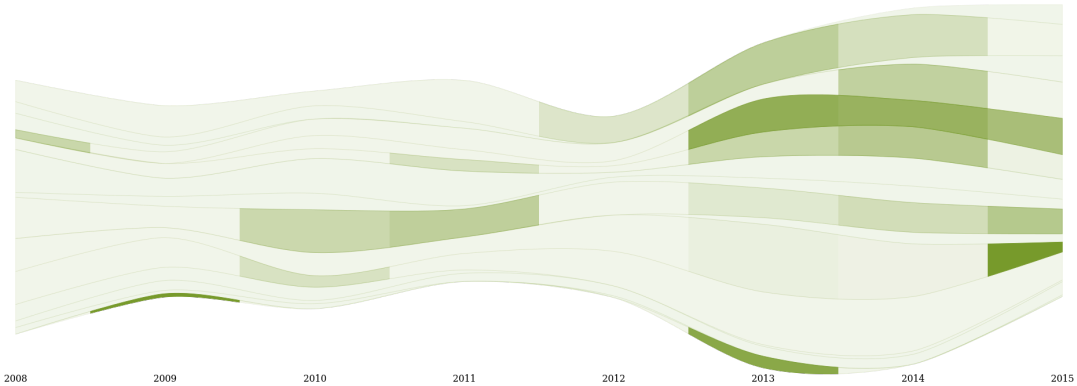


Abbildung 4.7.: Gradient. Alle Institute der Informatik wurden ausgewählt. Der Gradient berechnet das Verhältnis der interfakultären Kooperationen zum Gesamtwert einer jeden Schicht zu jedem Zeitpunkt. Der Gradient wurde normiert, um die Farben deutlicher hervorzuheben (da in diesem Fall $max(y_{part}) = 0.5$ gilt, wurde die Deckkraft aller Farben durch die Normierung verdoppelt).

Endpunkt B einer Kante ergibt sich entsprechend der beiden Formeln

$$A = P_a(t-1) + 0.5(P_a(t) - P_a(t-1)); B = P_b(t) + 0.5(P_b(t+1) - P_b(t)) \quad (4.3)$$

, wobei t der aktuell betrachtete Zeitschritt ist und P die definierten Punkte der kooperierenden Institute a und b sind. Bei der Nutzung von Cardinal Splines zur Berechnung des Streamgraphen müsste die Formel entsprechend angepasst werden, damit die Endpunkte der Kanten tatsächlich in den zugehörigen Schichten liegen. Für jede Kooperation entstehen zwei sich überkreuzende Kanten mit demselben Anstiegsbetrag, aber unterschiedlichen Vorzeichen, in der Darstellung.

Die Visualisierung hat dieselben Nachteile wie die aus Abbildung 3.8 mit weiteren Einschränkungen, weil die Kanten in einigen Bereichen noch enger aneinander gedrängt sind als zuvor und weil sich die Position der Institute über die Zeit verändert, wodurch es schwieriger ist, die Entwicklung ihrer Kanten nachzuvollziehen. Um dem entgegenzuwirken werden nur noch die Kanten der Schicht angezeigt, welche aktuell im Fokus des Nutzers liegt (und deren Tooltip angezeigt wird), wodurch die Überfüllung der Visualisierung deutlich nachlässt und dennoch die speziellen Kooperationsbeziehungen eines jeden Instituts analysiert werden können. Eine Beispielanwendung wird in Abbildung 4.9 dargestellt. Eine allgemeine Übersicht über das Netzwerk ist in diesem Fall natürlich nicht mehr gegeben.

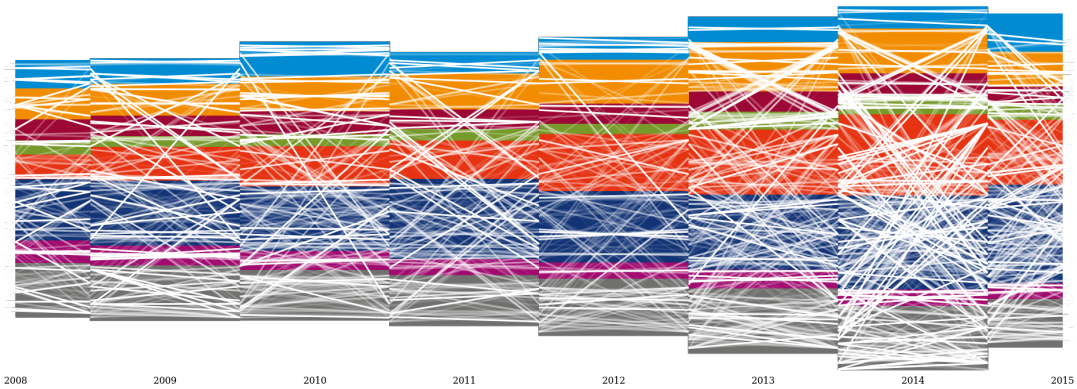


Abbildung 4.8.: Kooperationskanten. In jedem Zeitschritt wird eine Kante zwischen je zwei Instituten gezogen, welche in dem Jahr kooperierten. Die Endpunkte der Kante werden durch lineare Interpolation entsprechend der Formel 4.3 errechnet. Die Deckkraft der Kante wird ermittelt, indem der Wert einer jeden Kante durch den Maximalwert in allen Kanten einer Schicht geteilt wird.

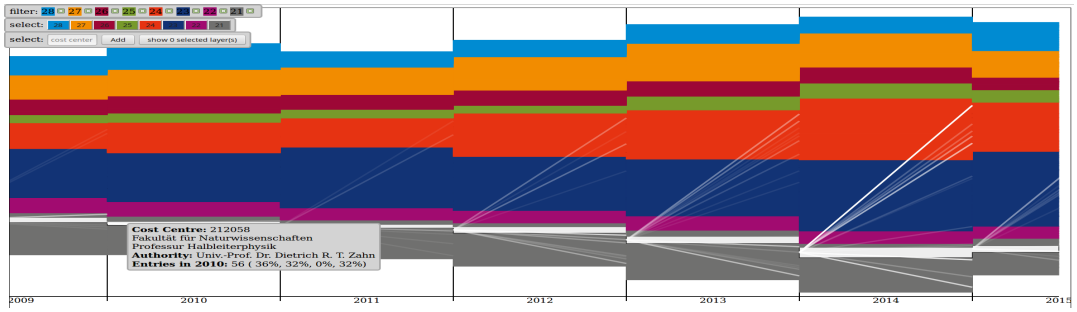


Abbildung 4.9.: Fokussierte Kooperationskanten. Die Kanten werden nur für das fokussierte, weiß eingefärbte, Institut *Professur Halbleiterphysik* angezeigt. Neben der zeitlichen Entwicklung der allgemeinen Kooperationen des Instituts durch die Dicke der Schicht ist nun auch erkennbar, mit welchen speziellen Instituten koopiert wurde und wie sich diese Kooperationen entwickelten. So lässt sich sagen, dass dieses Institut hauptsächlich in der eigenen Fakultät und mit der orangen Fakultät, sowie hin und wieder mit einigen blauen Instituten kooperierte. Dabei scheint von 2011 bis 2014 sowohl die Kooperation mit einem grauen (unterste Kante), als auch einem orangen (oberste Kante) Institut erhalten geblieben und stets gewachsen zu sein, was an der steigenden Deckkraft der Kanten ersichtlich ist.

5. Evaluation

5.1. Einleitung

Nachdem alle Funktionalitäten der implementierten Anwendung in Kapitel 4 beschrieben wurden, soll eine qualitative Evaluierung zeigen, ob die Visualisierung mit ihren Interaktionsmöglichkeiten für ihren Anwendungszweck geeignet ist, ob Probleme bei der Verwendung auftreten, oder zusätzliche Funktionen benötigt werden.

5.2. Vorgehen

An der Evaluation haben fünf freiwillige Probanden verschiedenen Geschlechts, Alters und Bildungsgrades teilgenommen. Alle Probanden waren den Umgang mit Computern gewohnt. Die Kenntnis über die zugrundeliegenden Daten hat derart variiert, dass ein Proband die allgemeinen Strukturen einer Universität nicht kannte, einer mit ihnen vertraut war, einer regelmäßig mit den Publikationsdaten selbst zu tun hat und zwei die Publikationsdaten zur Netzwerkanalyse [AB16] verwendet haben.

Die Probanden haben eine Einführung in die Zielstellungen und Informationsrepräsentation der Anwendung erhalten. Anschließend konnten sie alle Interaktionsmöglichkeiten selbst austesten, währenddessen ihre Funktionalität und ihr jeweiliger Einsatzzweck erklärt wurden. Den Probanden wurden weiterhin zehn Fragen gestellt, welche mithilfe der Anwendung beantwortet werden können und in Anhang A zu finden sind. Die Beantwortung der Fragen soll dem Probanden einen Einblick darin geben, welche Informationen aus der Visualisierung herausgelesen werden können und welche Optionen der Anwendung kombiniert werden können, um diese zu erhalten. Die Probanden wurden dazu ermuntert, ihre Gedanken bei der Benutzung des Programms laut auszusprechen, um auf eventuelle Probleme und Uneindeutigkeiten hinzuweisen. Weiterhin wurden sie bei der Bearbeitung der Aufgaben beobachtet und ihr Vorgehen analysiert. Kam ein Proband an einer Stelle nicht weiter, wurden erst richtungsweisende Hinweise gegeben und im Notfall einzelne Optionen vorgeschlagen, welche die Fragen beantwortbar machten. Schließlich wurde den Probanden Zeit gegeben, um ihre Meinung zu der Anwendung in Hinblick auf das Gefallen der Visualisierung, die Verwendbarkeit für die gestellten Aufgaben, Positives und Negatives zur Funktionalität und Bedienbarkeit der Anwendung zu äußern und Verbesserungsvorschläge zu liefern. Die drei Probanden, welche mit den Publikationsdaten vertraut waren, wurden weiterhin gefragt, ob sie sich vorstellen könnten, mit dieser Anwen-

dung zu arbeiten.

5.3. Auswertung

Ausgenommen vereinzelter Schwierigkeiten beim Verständnis der Fragestellungen ließen sich alle Aussagen der Probanden den drei Kategorien Visualisierung, Ergonomie und Erweiterungen zuordnen. Diese beschreiben die Art der Darstellung der Daten, die Umsetzung der Steuerbarkeit zur Anpassung der Darstellung und Vorschläge zur Verbesserung der Anwendung um sie entweder effizienter, oder für weitere Zwecke nutzen zu können. Die Aussagen wurden in den folgenden Abschnitten jeweils ihrer Kategorie zugeordnet und inhaltlich zusammengefasst.

5.3.1. Visualisierung

Der Streamgraph als Visualisierung für die Kooperationsdaten hat allen Probanden generell gut gefallen. Die Farbgebung ohne Gradienten wirkte auf einen Probanden „unprofessionell", wurde jedoch von anderen, welche bereits mit den Publikationsdaten zu tun hatten, als positiv empfunden, weil sie die Schichten direkt ihren Fakultäten zuordnen konnten.

Für alle Teilnehmer war es anfangs schwierig, die Funktionsweise des Gradienten und die dadurch übermittelten Informationen zu verstehen. Nach einer gewissen Einarbeitungszeit wurde das Prinzip klarer und er konnte gut verwendet werden, um Institute mit ausgewählten Kooperationsverhältnissen hervorzuheben. Einige Probanden wiesen darauf hin, dass der Gradient der Darstellung nur zuträglich ist, wenn die Anzahl der Schichten reduziert wurde, während er in der Anzeige aller Schichten mehr anzeigt, als vom Nutzer sinnvoll verarbeitet werden kann.

Die Probanden gaben weiterhin an, dass mit der präsentierten Darstellung der Vergleich zwischen ausgewählten Instituten sehr gut herstellbar ist und im Modus *expand* auch prozentuale Verhältnisse sehr gut abgelesen werden können, wobei dieser insbesondere von einer Skale profitieren würde. Besonders positiv viel die Bewertung der Darstellung zeitlicher Entwicklungen aus, welche von den Teilnehmern sehr ausführlich und detailliert beschrieben werden konnten.

5.3.2. Ergonomie

Während die Visualisierung zur Bewältigung der Aufgaben geeignet ist, wurden bei der Wahl der bereitgestellten Menüs zur Anpassung der Optionen einige Probleme festgestellt. In erster Linie fehlte den Probanden ein *Reset* Knopf, welcher die komplette Anwendung wieder in ihren Urzustand versetzt. In eine ähnliche Richtung geht der Wunsch nach einem *Zurück*-Knopf, welcher die letzte Aktion ungeschehen macht. In einem Einzelfall hat ein Proband aus Versehen im *expand*-Modus auf eine Schicht

geklickt, wodurch diese in Form eines Rechteckes die komplette Anwendung einnahm und der Proband nicht wusste, wie er dies rückgängig machen könne. In solchen Fällen würden die beiden vorgeschlagenen Optionen einen Ausweg bieten.

Bei der gegebenen Auswahl an Sortierfunktionen hat die Namensgebung dazu geführt, dass alle Probanden nicht genau wussten, wonach die Schichten sortiert werden. Insbesondere die in Kapitel 4.4 dargestellte Sortierung innerhalb der Fakultät veranlasste alle Probanden zu glauben, dass die Fakultäten nach ihren Werten sortiert würden und nicht die Schichten innerhalb der Fakultäten. Die angesprochene Sortierfunktion sollte derart verändert werden, dass zusätzlich zu den Schichten in den Fakultäten auch die Fakultäten selbst sortiert werden, weil dann auch ein objektiver Vergleich zwischen Fakultäten möglich ist, wenn diese ungefähr gleich große Werte haben.

Den Anwendern fiel es zum Teil schwer, die Funktionen des Filter- und Selektionsmechanismus auseinanderzuhalten, wodurch sie hin und wieder das falsche auswählten und sich über das Ergebnis wunderten. Ebenso kann aktuell eine einzelne Schicht, nachdem sie für die Fokussierung ausgewählt wurde, nicht deselektiert oder aus dem Fokus entfernt werden, da ein Klick auf die fokussierte Darstellung wieder zur globalen Ansicht führt. Über die Umsetzung der Filter- und Selektionsfunktion sollte noch einmal nachgedacht werden, damit insbesondere auch innerhalb der fokussierten Darstellung erneut Schichten ausgewählt werden können, um den Fokus zu erhöhen.

Obwohl das zeitliche Scrollen der Anwendung sowohl mithilfe des Scrollbalkens, als auch des Mausrades möglich war, versuchten die Probanden zuerst, ähnlich der Gestenfunktion von berührungsempfindlichen Bildschirmen, mit der Maus an eine Position zu klicken und die Anwendung zur Seite zu schieben.

Bei Menüs mit mehreren Ankreuzfeldern sollten Knöpfe zum An- und Abhaken aller Felder gleichzeitig bereitgestellt werden, um die gezielte Darstellung schneller zu erreichen.

5.3.3. Erweiterungen

Durch die ausführliche Nutzung der Anwendung kamen den Probanden verschiedene Ideen zu deren Verbesserung. Neben den bereits gegebenen Sortierfunktionen zur Hervorhebung der höchsten Werte in einem Jahr wurde weiterhin die Sortierung nach den höchsten Durchschnittswerten in einem bestimmten Zeitraum vorgeschlagen, sowie auch die Sortierung nach den Gradientenwerten, wenn sie zur Anzeige verwendet werden.

Neben den bereits besprochenen Knöpfen zum Zurücksetzen auf Standard und dem Zurücksetzen der letzten Aktion, wurde weiterhin ein System von Reitern (*Tabs*) vorgeschlagen, welches die Speicherung aller gewählten Optionen vornimmt und eine neue Ansicht zur Bearbeitung bereitstellt, sodass anschließend sowohl mehrere

Ansichten miteinander verglichen werden können, als auch auf zuvor abgespeicherte Selektionen zurückgegriffen werden kann.

Zusätzlich zur Speicherung von gewählten Optionen ist auch die Bereitstellung einiger häufig genutzter Optionskombinationen gewünscht, welche besonders gut geeignet sind, um spezielle Informationen in der Visualisierung hervorzuheben. Diese könnten sowohl die Einarbeitungszeit der Nutzer verkürzen, als auch erfahrene Nutzer schneller zur gewünschten Darstellung bringen, indem erst eine der möglichen Kombinationen gewählt und anschließend nur noch in einzelnen Bereichen angepasst wird.

Für die aktive Nutzung der Anwendung in der Bibliothek und Verwaltung sollte die Berechnung der Kooperationswerte nach der Dokumentart der Publikationen und nach ihrer öffentlichen Verfügbarkeit gefiltert werden können.

Zwar ist die Datenbasis aktuell nicht dafür ausgelegt, jedoch gibt es zukünftig Anstrengungen, die Publikationskosten einer jeden Veröffentlichung in die Datenbank einzutragen. Sollte dies der Fall sein, wäre auch eine Analyse der jeweiligen Kosten und eventuell erwirtschafteten Mittel interessant.

5.4. Zusammenfassung

Die Evaluation hat ergeben, dass die gewählte Visualisierung durchaus dafür geeignet ist, Institute in Hinblick auf ihre Kooperationen zu untersuchen und zu vergleichen. Die unterschiedlichen Optionen bieten die Möglichkeit, Einblick in verschiedenartige, für den Nutzer interessante, Informationen zu erhalten, allerdings sorgt die Anzahl an Parametern für eine längere Einarbeitungszeit in die Anwendung. Hat sich ein Nutzer an die Funktionalitäten gewöhnt, ist die Anwendung bereits jetzt für Wissenschaftler interessant, um zu schauen, wo sie im Vergleich zu anderen Instituten an ihrer Universität stehen und wie ihre Entwicklung im allgemeinen Vergleich einzuschätzen ist.

Um die Anwendung auch für Personen interessant zu machen, welche regelmäßig mit den Publikationsdaten arbeiten, sollte sie um die Filterung der Publikationen nach ihrer öffentlichen Verfügbarkeit und ihrer Dokumentart erweitert werden, um beispielsweise nur die Kooperationen bei referierten Veröffentlichungen zu betrachten. Weiterhin sollte allgemein die Ergonomie der Anwendung verbessert werden, um sie in den Produktiveinsatz zu bringen.

6. Anwendungsbeispiele

Nachdem die vorhanden Interaktionsmöglichkeiten der Anwendung in Kapitel 4 vorgestellt wurden und ihr Nutzen in Kapitel 5 bestätigt wurde, werden in diesem Kapitel einige Anwendungsbeispiele vorgestellt, welche zur Erläuterung der Optionskombinationen dienen und zeigen, welche Art von Informationen aus der Visualisierung gelesen werden können. Dabei wird auch bestätigt, dass die anfangs gesetzten Zielstellungen dieser Arbeit durch die Anwendung erfüllt worden sind.

Abbildung 4.3 zeigte bereits, wie Schichten nach ihren Werten in einzelnen Jahren sortiert werden können. Wählt man zur Anzeige nur Kooperationen mit anderen Fakultäten aus und sortiert nach den Werten im aktuellen Jahr 2015, lässt sich sehr gut erkennen, welche Institute am stärksten interdisziplinär kooperieren und welcher Fakultät diese angehören. Die Kombination wird in Abbildung 6.1 dargestellt.

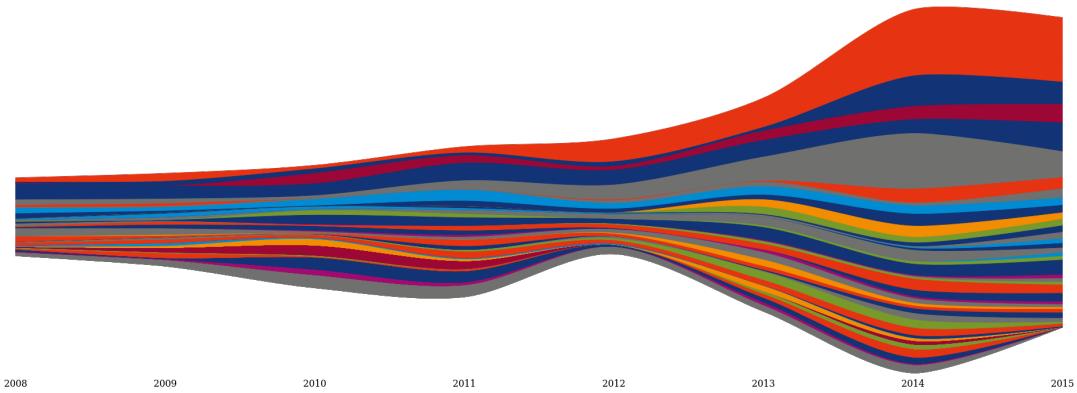


Abbildung 6.1.: Die Auswahl interfakultärer Kooperationen ermöglicht in Kombination mit der Sortierung nach Werten im Jahr 2015 die Bestimmung der Fakultäten, welche aktuell am meisten untereinander kooperieren. Die leuchtorange, dunkelblaue und graue Fakultät führen das Feld recht deutlich an, wobei die leuchtorange bereits seit 2012 der Spitzenreiter ist.

Die Sortierung ist nicht nur zum Auffinden von Maximalwerten gut, sondern kann auch Institute zum Vorschein bringen, welche nach einer Neugründung eine gute Entwicklung hingelegt haben. Abbildung 6.2 zeigt, wie die Sortierung nach dem Wert

in einem frühen Jahr in späteren Jahren Institute mit starker Steigerung hervorhebt. Diese Betrachtung kann erneut auf unterschiedlicher Datenlage erfolgen, indem zum Beispiel nach starken Steigerungen in externen Kooperationen gesucht wird.

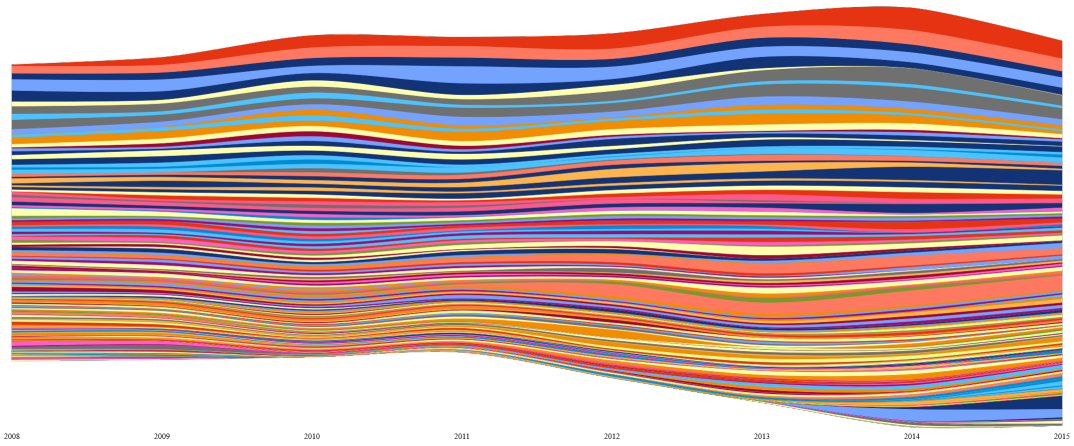
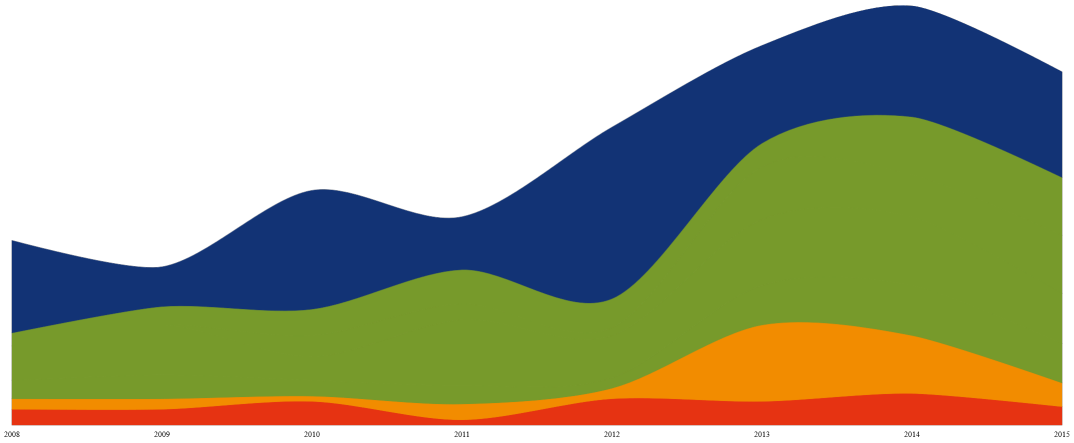


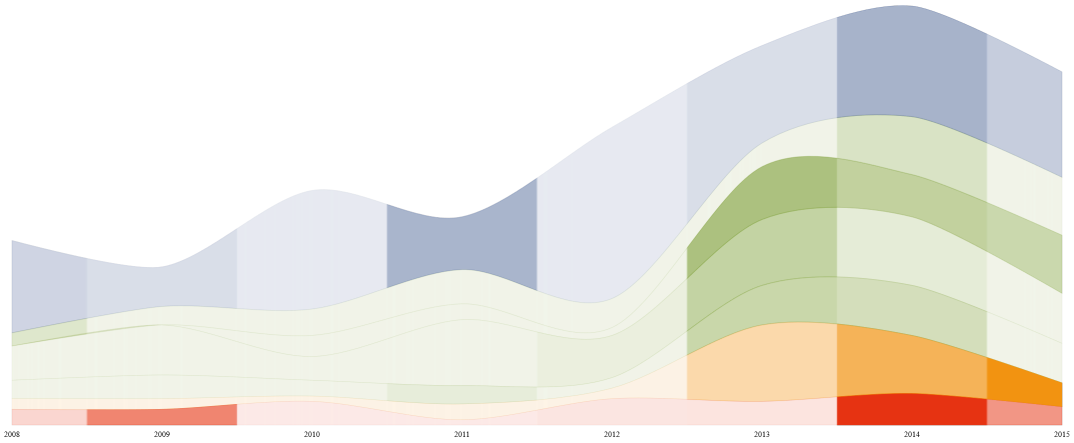
Abbildung 6.2.: Die Sortierung nach den Werten im Jahr 2010 sammelt niedrige Werte in diesem Jahr in der unteren Hälfte des Streamgraphen. Betrachtet man nun das aktuelle Jahr 2015 in der unteren Hälfte, zeigen sich Institute, welche eine starke Steigerung erfahren haben. In diesem Fall betrifft dies zwei blaue und ein oranges Institut. Alternierende Farben wurden aktiviert, um die beiden blauen Institute auseinanderhalten zu können.

An der TU Chemnitz existiert seit 2011 das *Kompetenzzentrum Virtual Humans*, dessen zugehörige Professuren in Abbildung 6.3 ausgewählt wurden. Es zeigt sich, dass die beteiligten Professuren seit der Entstehung des Kompetenzzentrums einen starken Zuwachs erhalten haben, welcher insbesondere in der Fakultät für Informatik (grün) festzustellen ist.

Betrachtet man eine komplette Fakultät wie in Abbildung 6.4 so können sowohl allgemeine, als auch spezielle Informationen gleichzeitig analysiert werden. Einerseits fällt auf, dass die Gesamtwerte seit 2012 stetig zugenommen haben, andererseits zeigt sich, dass die Steigerung in den Jahren 2014 und 2015 hauptsächlich auf zwei neue Institute zurückgeführt werden kann, welche die Plätze zwei und vier in der aktuellen Rangfolge belegen. Möchte man herausfinden, woher diese beiden Institute ihre Steigerung hauptsächlich erlangt haben, können verschiedene Gradienten getestet werden, bis schließlich eine Auffälligkeit bei den externen Kooperationen hervortritt.

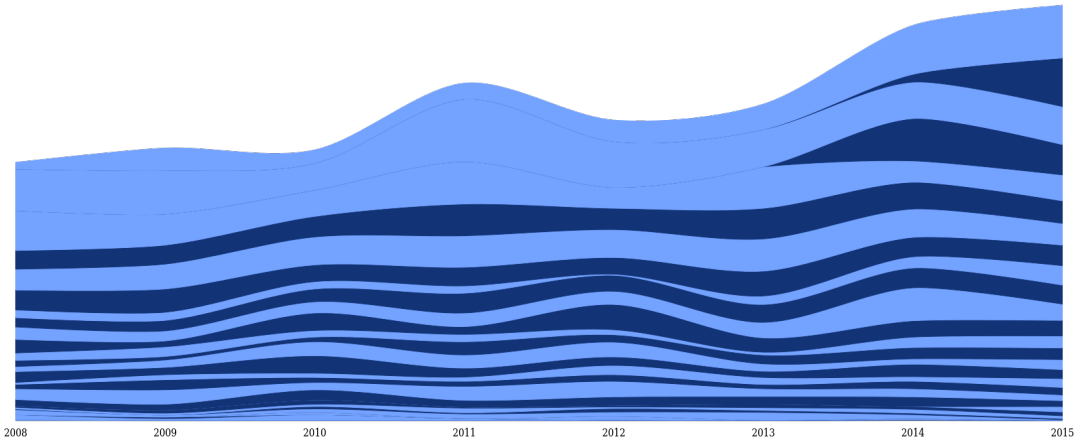


(a) Die ausgewählten Institute zeigen eine starke Wertsteigerung seit der Gründung des Kompetenzzentrums im Jahr 2011.

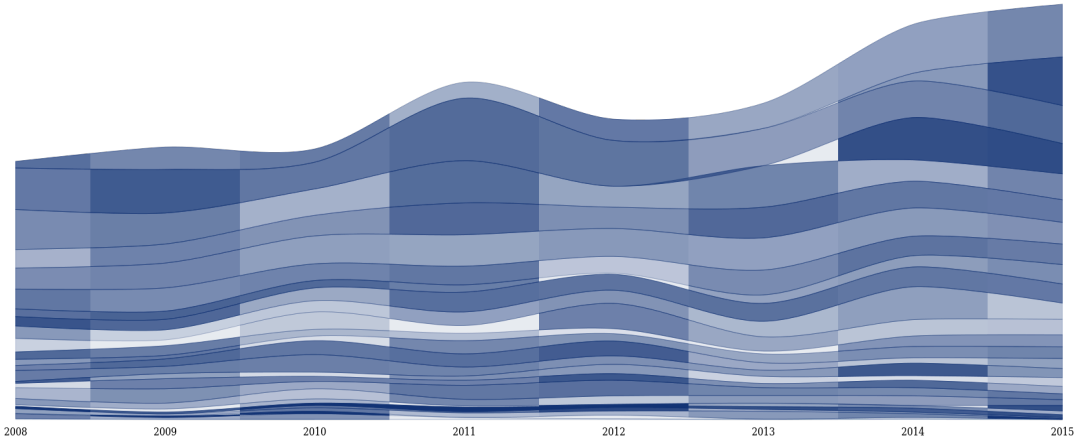


(b) Der Gradient hebt interdisziplinäre Kooperationen hervor und zeigt, dass der starke Anstieg der Werte der Professuren hauptsächlich auf diese Zusammenarbeit zurückzuführen ist.

Abbildung 6.3.: Kompetenzzentrum Virtual Humans



(a) Die Alternierung der Farben dient der Unterscheidung der Schichten. Zwei neue Institute sind für die Steigerung in den Jahren 2014 und 2015 verantwortlich.



(b) Der Gradient zeigt den Anteil externer Kooperationen (alternierende Farben wurden deaktiviert) und wurde normiert, um die Deckkraft der Farben zu erhöhen. Die beiden neuen Institute haben verhältnismäßig mehr mit externen Mitarbeitern kooperiert als alle anderen Institute der Fakultät (mit Ausnahme der untersten Institute).

Abbildung 6.4.: Fakultät für Maschinenbau, sortiert nach dem Wert im Jahr 2015.

7. Zusammenfassung und Ausblick

In der vorliegenden Arbeit wurde eine Anwendung auf Basis des Streamgraphen vorgestellt, welche Nutzern ermöglicht, die zeitliche Entwicklung von Kooperationen zwischen Instituten der TU Chemnitz zu untersuchen und dabei sowohl verschiedene Institute, als auch Fakultäten miteinander zu vergleichen. Dieses Kapitel dient der Zusammenfassung der geleisteten Arbeit und der Funktionsweise der entstandenen Anwendung, zeigt anhand der Evaluierung und Anwendungsbeispiele aus den Kapiteln 5 und 6, dass die anfangs gestellten Zielstellungen von der gewählten Visualisierung erfüllt werden und gibt einen Ausblick in die zukünftig mögliche Arbeit zur Verbesserung der Anwendung.

7.1. Zusammenfassung

Die Datenanalyse in Kapitel 2 zeigte auf, dass die Aufstellung eines Kooperationsnetzwerkes auf Basis der gegebenen Daten ausschließlich anhand der Mehrautorenschaft der Publikationen stattfinden kann. Es wurde gezeigt, dass die Darstellung des Netzwerkes in Form eines Graphen dazu in der Lage ist, die Beziehungsstruktur zwischen Fakultäten, und mit Einschränkungen auch zwischen Instituten, aufzuzeigen. Für die Erweiterung um eine zeitliche Komponente ist sie jedoch nicht geeignet, weshalb mit dem Streamgraphen eine Darstellung gewählt wurde, welche von Grund auf für die Verarbeitung zeitlich bedingter Daten angedacht ist.

Der Streamgraph erstellt für jedes Institut einen Punkt mit Angabe des Jahres und dem erreichten Kooperationswert. Durch Interpolation der Punkte eines Instituts entstehen so viele Polygonzüge, wie es Institute gibt, welche anschließend durch Aufsummierung der Werte am jeweiligen Zeitpunkt auf einer Grundlinie übereinandergelegt werden. Der Bereich zwischen zwei Polygonzügen wird jeweils als Schicht dieses Instituts verwendet und in der Farbe der zugehörigen Fakultät eingefärbt. Auf dieser Basis, und mithilfe von Sortierfunktionen, können Institute im allgemeinen Vergleich zur Gesamtheit der publizierenden Institute und zu Fakultäten gesetzt werden. Die Selektion und Filterung von Instituten fokussiert den Blick des Nutzers auf wesentliche Schichten und ermöglicht detailliertere Aussagen zur Entwicklung und dem Verhältnis ausgewählter Institute. Die Nutzung des Gradienten kann den Blick zusätzlich auf spezielle Verhältnisse zwischen Kooperationskategorien hinweisen, indem die Farben einzelner Institute und Zeitbereiche transparenter dargestellt werden und somit der Blick auf die weniger transparenten Bereiche fällt.

Die Evaluation bekräftigt die Aussage, dass die verwendete Visualisierung mithilfe zusätzlicher Funktionalitäten und verschiedener Interaktionsmöglichkeiten sowohl allgemeine, als auch spezifische Aussagen über die Kooperationen von Instituten ermöglicht, obwohl die Netzstruktur selbst nicht mehr erkennbar ist. Zusätzlich wurde bestätigt, dass die zeitliche Entwicklung der dargestellten Schichten sehr gut repräsentiert wird und detailliert beschrieben werden kann. Viele der im Ausblick folgenden Ideen zur zukünftigen Ausgestaltung der Anwendung entstammen den Auswertungen der Evaluierung.

7.2. Ausblick

Zwar ist die hier vorgestellte Anwendung bereits jetzt nützlich, um Wissenschaftlern der TU Chemnitz einen Einblick in den Stand ihres Instituts zu gewähren, jedoch können einige Funktionen angepasst und Erweiterungen hinzugefügt werden, um die Anwendung weiter zu verbessern und ihren Einsatz für verwaltungstechnische Ansprüche zu legitimieren.

Um die Attraktivität der Anwendung für die Bibliothek und Verwaltung der TU Chemnitz zu steigern, wird die Filterung der Daten nach Dokumentart und öffentlicher Zugänglichkeit benötigt. Da mit dieser Funktionalität mehr Nutzer angesprochen werden, wird ihr eine hohe Priorität zugesprochen. Weiterhin zeigte die Evaluation einige Probleme in der Nutzung der in der Anwendung verwendeten Bedienelemente auf, welche angepasst und anschließend neu evaluiert werden sollten, um den Produktiveinsatz zu ermöglichen. Auch zusätzliche Bedienelemente wie die Speicherung einer Optionskombination und die Führung einer Historie der zuletzt getätigten Aktionen, um deren Zurücksetzung zu ermöglichen, können die Ergonomie der Anwendung deutlich verbessern.

Einer der grundlegendsten Bausteine der Visualisierung sind die Daten, wobei sich das, was dargestellt wird von dem unterscheiden kann, was ein Nutzer meint zu sehen. Da die Anwendung auf der Basis von Publikationen arbeitet, wird die Annahme, dass konkrete Anzahlen von Publikationen in den Jahren angezeigt werden, recht häufig vorkommen. Tatsächlich wird jedoch für jedes Institut und jedes Jahr ein Kooperationswert berechnet, welcher in den meisten Fällen höher ausfällt als die Anzahl der getätigten Publikationen und das Verhältnis des berechneten Wertes zum konkreten Wert stark schwanken kann. Es sollte evaluiert werden, wie häufig die beschriebene Annahme getroffen wird und ob die in Kapitel 2.7 beschriebene partielle Zählung trotz geringerer Werte in der Lage ist, stark kooperierende Institute hervorzuheben.

Obwohl der Streamgraph bereits in vielen Anwendungen Verwendung fand, wurde seine Erweiterbarkeit erst wenig erforscht. Zusätzliche Elemente wie der hier vorgestellte Gradient und die Kooperationskanten können den Einsatz dieser Visualisierung in neuen Anwendungsfällen ermöglichen und sollten exploriert werden.

Literaturverzeichnis

[AB16] Carolin Ahnert und Martin Bauschmann: *Interdisziplinäres Publikationsnetzwerk der TU Chemnitz*, 2016, unpublished thesis.

[ABA03] Lada Adamic, Orkut Buyukkokten und Eytan Adar: *A social network caught in the web*, *First monday*, Bd. 8(6), 2003.

[BC03] Ulrik Brandes und Steven R Corman: *Visual unrolling of network evolution and the analysis of dynamic discourse†*, *Information Visualization*, Bd. 2(1):S. 40–50, 2003.

[Ber99] François Bertault: *A force-directed algorithm that preserves edge crossing properties*, in *International Symposium on Graph Drawing*, S. 351–358, Springer, 1999.

[BH86] Josh Barnes und Piet Hut: *A hierarchical O (N log N) force-calculation algorithm*, *nature*, Bd. 324(6096):S. 446–449, 1986.

[BHJ+09] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy *et al.*: *Gephi: an open source software for exploring and manipulating networks.*, *ICWSM*, Bd. 8:S. 361–362, 2009.

[BOH11] Michael Bostock, Vadim Ogievetsky und Jeffrey Heer: *D^3 data-driven documents*, *IEEE transactions on visualization and computer graphics*, Bd. 17(12):S. 2301–2309, 2011.

[BP11] Ulrik Brandes und Christian Pich: *Explorative visualization of citation patterns in social network research*, *Journal of social structure*, Bd. 12(8):S. 1–19, 2011.

[BW08] Lee Byron und Martin Wattenberg: *Stacked graphs–geometry & aesthetics*, *IEEE transactions on visualization and computer graphics*, Bd. 14(6):S. 1245–1252, 2008.

[coo14] *Handbuch zum Corporate Design der Technischen Universität Chemnitz*, `https://www.tu-chemnitz.de/uk/corporate_design/handbuch/handout_web.pdf`, 2014, [Online; accessed 15-September-2016].

[CP96] Michael K Coleman und D Stott Parker: *Aesthetics-based Graph Layout for Human Consumption*, *Software: Practice and Experience*, Bd. 26(12):S. 1415–1438, 1996.

[DBH16] Marco Di Bartolomeo und Yifan Hu: *There is More to Streamgraphs than Movies: Better Aesthetics via Ordering and Lassoing*, in *Computer Graphics Forum*, Bd. 35, S. 341–350, Wiley Online Library, 2016.

[Dwy09] Tim Dwyer: *Scalable, versatile and simple constrained graph layout*, in *Computer Graphics Forum*, Bd. 28, S. 991–998, Wiley Online Library, 2009.

[EAD84] P. EADES: *A Heuristics for Graph Drawing*, *Congressus Numerantium*, Bd. 42:S. 146–160, 1984.
URL `http://ci.nii.ac.jp/naid/10000075358/en/`

[FLM94] Arne Frick, Andreas Ludwig und Heiko Mehldau: *A fast adaptive layout algorithm for undirected graphs (extended abstract and system demonstration)*, in *International Symposium on Graph Drawing*, S. 388–403, Springer, 1994.

[FR91] Thomas MJ Fruchterman und Edward M Reingold: *Graph drawing by force-directed placement*, *Software: Practice and experience*, Bd. 21(11):S. 1129–1164, 1991.

[Fre00] Linton C Freeman: *Visualizing social networks*, *Journal of social structure*, Bd. 1(1):S. 4, 2000.
URL `http://www.cmu.edu/joss/content/articles/volume1/Freeman.html`

[gal] *Galaxiensimulation mit dem Barnes-Hut Algorithmus*, `http://beltoforion.de/article.php?a=barnes-hut-galaxiensimulation&hl=de&s=idBarnesHut#idBarnesHut`, [Online; accessed 12-September-2016].

[GR87] Leslie Greengard und Vladimir Rokhlin: *A fast algorithm for particle simulations*, *Journal of computational physics*, Bd. 73(2):S. 325–348, 1987.

[HB05] Jeffrey Heer und Danah Boyd: *Vizster: Visualizing online social networks*, in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, S. 32–39, IEEE, 2005.

[HHN00] Susan Havre, Beth Hetzler und Lucy Nowell: *ThemeRiver: Visualizing theme changes over time*, in *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, S. 115–123, IEEE, 2000.

[int11] *Interdisziplinäres Kompetenzzentrum*, `https://www.tu-chemnitz.de/forschung/virtual_humans/`, 2011, [Online; accessed 12-September-2016].

[ITK10] Masahiko Itoh, Masashi Toyoda und Masaru Kitsuregawa: *An interactive visualization framework for time-series of web graphs in a 3D environment*, in *2010 14th International Conference Information Visualisation*, S. 54–60, IEEE, 2010.

[JVHB14] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann und Mathieu Bastian: *ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software*, *PloS one*, Bd. 9(6):S. e98679, 2014.

[KG06] Gautam Kumar und Michael Garland: *Visual exploration of complex time-varying graphs*, *IEEE Transactions on Visualization and Computer Graphics*, Bd. 12(5):S. 805–812, 2006.

[Kob04] Stephen G Kobourov: *Force-directed drawing algorithms*, 2004.

[MGF12] Justin Matejka, Tovi Grossman und George Fitzmaurice: *Citeology: visualizing paper genealogy*, in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, S. 181–190, ACM, 2012.

[New04] Mark EJ Newman: *Coauthorship networks and patterns of scientific collaboration*, *Proceedings of the national academy of sciences*, Bd. 101(suppl 1):S. 5200–5205, 2004.

[Noa07] Andreas Noack: *Energy Models for Graph Clustering.*, *J. Graph Algorithms Appl.*, Bd. 11(2):S. 453–480, 2007.

[PRWvE16] Antonio Perianes-Rodriguez, Ludo Waltman und Nees Jan van Eck: *Constructing bibliometric networks: A comparison between full and fractional counting*, *arXiv preprint arXiv:1607.02452*, 2016.

[RFF+08] George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee und John Stasko: *Effectiveness of animation in trend visualization*, *IEEE Transactions on Visualization and Computer Graphics*, Bd. 14(6):S. 1325–1332, 2008.

[Stu06] Universitätsbibliothek Stuttgart: *Technische Dokumentation zu OPUS Version 3.0*, `https://www.bibliothek.tu-chemnitz.de/uni_biblio/API/opus-techdoku-3_0.pdf`, 2006, [Online; accessed 15-September-2016].

[vEW14a] Nees Jan van Eck und Ludo Waltman: *CitNetExplorer: A new software tool for analyzing and visualizing citation networks*, *Journal of Informetrics*, Bd. 8(4):S. 802–823, 2014.

[vEW14b] Nees Jan van Eck und Ludo Waltman: *Visualizing bibliometric networks*, in *Measuring scholarly impact*, S. 285–320, Springer, 2014.

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig angefertigt, nicht anderweitig zu Prüfungszwecken vorgelegt und keine anderen als die angegebenen Hilfsmittel verwendet habe. Sämtliche wissentlich verwendete Textausschnitte, Zitate oder Inhalte anderer Verfasser wurden ausdrücklich als solche gekennzeichnet.

Chemnitz, den 22. September 2016

Fabian Bolte

Anhang A.

Fragestellungen der Evaluation

1.) Welche Professur hat im Jahr 2012 die meisten Einträge?
2.) Welche Fakultät hat im Jahr 2012 die wenigsten Einträge?
3.) Welches Jahr von 2010 bis 2015 hat insgesamt die wenigsten Einträge?
4.) Welche Professur hat in der Philosophischen Fakultät den größten Einfluss?

Die Fragen 5 und 6 können mithilfe des Gradienten beantwortet werden.

5.) Welche Professur hat von 2011 bis 2013 prozentual mehr innerhalb der eigenen Fakultät kooperiert, als alle anderen Professuren?
6.) Welche Professur der Informatik hat in den Jahren 2013 bis 2015 jeweils bei über 30% ihrer Veröffentlichungen mit anderen Fakultäten kooperiert?

Wählen Sie für die Fragen 7-10 die beiden Kostenstellen 231533 und 249000 aus.

7.) Wie hoch ist der geschätzte, prozentuale Anteil der Kostenstelle 231533 von den aktuell angezeigten Einträgen im Jahr 2010?
8.) Beschreiben Sie die zeitliche Entwicklung der beiden Kostenstellen in Worten.
9.) Welche der beiden Kostenstellen würden Sie als kooperativer beschreiben und warum?
10.) Wieviele Kooperationen mit anderen Fakultäten hat die Kostenstelle 249000 von 2012 bis 2014?

Anhang B.

USB-Stick

Zusammen mit dieser Arbeit wird ein USB-Stick eingereicht. Dieser beinhaltet neben der Masterarbeit im PDF-Format drei Ordner namens *application*, *pics* und *papers*.

- *application* beinhaltet den aktuellen Stand der Anwendung, welche durch das Öffnen der Datei *index.html* im Browser gestartet werden kann
- *pics* beinhaltet alle in der Masterarbeit verwendeten Bilder
- *papers* beinhaltet die meisten der in der Masterarbeit zitierten Publikationen